MASTER EN INGÉNIEUR DE GESTION À FINALITÉ SPÉCIALISÉE EN ANALYTICS & DIGITAL BUSINESS

Application de la théorie des valeurs extrêmes dans le contexte de la gestion de l'information

Mignon, Pauline

Award date:
2021

Awarding institution:
Universite de Namur

Application de la théorie des valeurs extrêmes dans le contexte de la gestion de l'information

MIGNON Pauline

Directeur : Prof. A. Kiriliouk

Mémoire présenté en vue de l'obtention du titre de

Master 120 - Ingénieur de Gestion
Finalité Spécialisée en Analytics & Digital Business

ANNÉE ACADÉMIQUE : 2020-2021

Université de Namur, ASBL
Faculté des Sciences économiques, sociales et de gestion – Département des Sciences de gestion

Rempart de la Vierge 8, B-5000 Namur, Belgique, Tel. +32 [0]81 72 49 58/48 41

Remerciements

Je tiens à remercier infiniment ma directrice de mémoire, Anna Kiriliouk, qui m'a tant aidée, supportée et soutenue tout au long de la réalisation de mon mémoire. Elle a toujours été présente à tout moment. Je me sens chanceuse de l'avoir eue en tant que promotrice. Ensuite, je tiens à faire hommage à tous les professeurs qui, de par leur enseignement, m'ont permis d'arriver au terme de mes études et de présenter ce mémoire. Puis, je tiens à remercier mes parents et mon frère qui ont été d'un soutien sans faille et m'ont encouragé sans relâche durant toute cette période ainsi que ma famille et mes amis qui de près ou de loin m'ont supporté. Sans toutes ces personnes, ce mémoire aurait été plus difficile à réaliser.

Résumé

De nos jours, le trafic Internet ne cesse de croître chaque année en raison de l'omniprésence de plus en plus importante d'Internet dans nos vies. À l'heure actuelle, presque toutes les entreprises possèdent un site Internet et gèrent la majeure partie de leurs opérations via Internet. Il est donc crucial pour les entreprises de savoir gérer au mieux leur site et/ou leur système. Dès lors, ce travail se base sur la théorie des valeurs extrêmes du domaine des statistiques pour fournir aux entreprises une méthode leur permettant soit de déterminer le flux, c'est-à-dire le nombre de connexions, qu'elles doivent être capables de gérer pour un risque donné, soit de connaître le risque encouru d'un éventuel arrêt de leur site et/ou de leur système en fonction de la capacité de leur site et/ou système.

Summary

Nowadays, Internet traffic continues to grow every year due to the ever-increasing pervasiveness of the Internet in our lives. Nowadays, almost every company has a website and manages most of its operations via Internet. It is therefore crucial for companies to know how to best manage their site and/or their system. This work uses extreme value theory to provide companies with a method to either determine the flow, i.e. the number of connections, that they should be able to handle for a given risk, or to know the risk incurred by a possible shutdown of their site and/or their system according to the capacity of their site and/or system.

Table des matières

Application de la théorie des valeurs extrêmes dans le contexte de la gestion de l'information

Mignon Pauline

Juin 2021

1. Introduction

À l'heure actuelle, le trafic Internet ne cesse d'augmenter d'année en année. Ce dernier devient dès lors de plus en plus important à considérer. De nos jours, chaque entreprise possède un site Internet. De manière générale, les entreprises sont capables d'estimer de manière correcte leur nombre moyen de visiteurs. Cependant, il est plus difficile pour elles d'estimer leur nombre de visiteurs lors des «événements rares». L'évaluation approximative pour les événements rares cause très souvent des pannes (crashs) de sites Internet et/ou de systèmes et engendre des conséquences négatives d'un point de vue opérationnel et économique. En raison de ces pannes, les entreprises doivent faire face à des pertes de ventes potentielles ainsi qu'une perte de visibilité. Une maintenance est donc nécessaire afin de remettre le site et/ou les systèmes en marche et génère un coût important pour les entreprises. Il devient dès lors primordial pour les firmes de fixer un seuil pouvant supporter le flux entrant sur leur site Internet, c'est-à-dire un nombre fixe de visiteurs. Ce seuil correspond à la capacité du site Internet d'une entreprise. Si ce dernier est dépassé, l'entreprise fera face à un arrêt de son site car elle ne sera pas capable de gérer un flux de cette taille. Ce seuil est directement lié au risque accepté par une entreprise pour un arrêt de son site, c'est-à-dire à la probabilité correspondante de dépasser le seuil.

Sur base de ces informations, ce travail va employer la théorie des valeurs extrêmes afin de fournir une méthode rendant possible la gestion de cette problématique. La théorie des valeurs extrêmes est capable de prédire des événements rares. Cette technique permettra à une entreprise d'établir un seuil tel que la probabilité d'une panne (= le risque encouru) est fixée à une certaine (petite) valeur. Il sera également possible de savoir dans quelle mesure les pannes apparaitront dans le futur, c'est-à-dire la fréquence à laquelle elles se produiront. Les entreprises pourront dès lors sur base de ces informations récoltées décider si elles désirent prendre le risque que leur site encoure une panne ou si elles préfèrent prendre leur précaution et augmenter la puissance de leur site Internet, c'est-à-dire augmenter leurs ressources. Cela dépendra évidemment de la nature de l'activité de l'entreprise en question.

La question de recherche est donc la suivante:

Comment une entreprise peut-elle prédire une panne de son site Internet et/ou de ses systèmes à court et long terme sur base de ses données existantes en vue de maximiser la gestion de son site Internet et/ou de ses systèmes?

Par conséquent, deux scénarios sont envisagés.

Scénario 1: étant donné un flux, quelle est la probabilité de le dépasser? De par ce scénario, il serait possible d'estimer la probabilité d'avoir x visiteurs ou plus dans une période donnée, où x peut être supérieur au maximum de visiteurs observé dans le passé. Une entreprise peut savoir à partir de quel flux cela devient problématique pour elle. Cependant, elle ne sait pas dans quelle mesure ce flux peut apparaitre. Ce scénario permettra donc à une entreprise de déterminer le risque avec lequel un certain flux se produira.

Scénario 2: étant donné un risque, quel est le flux potentiel que l'on s'attend à atteindre ou à dépasser? Cela veut dire, qu'étant donné une probabilité p, une entreprise pourrait calculer le nombre maximal de visiteurs que le site doit être capable de gérer. La probabilité p sera petite et considérée par les entreprises comme le 'risque acceptable'. Une entreprise peut ne pas savoir à partir de quel flux cela devient problématique pour elle. Cependant, elle veut justement savoir à partir de quel flux elle entrera dans une situation aussi rare qu'elle apparaît tous les x ans. Ce scénario va permettre de donner cette valeur.

Ce travail va dès lors proposer une méthode en vue de trouver le niveau idéal de flux accepté pour une entreprise. Cela permettra aux entreprises de dépenser le montant adéquat à la gestion de son site Internet et/ou de ses systèmes. À partir des informations obtenues, une entreprise sera en mesure de décider si, oui ou non, elle doit changer le budget alloué à cette fin. Par conséquent, nous allons montrer comment les entreprises pourraient gérer au mieux leurs données et leur site Internet et/ou leurs systèmes en vue d'éviter un arrêt complet de ce(s) dernier(s).

Dans la section 2, nous allons passer en revue certaines études de la littérature ayant été effectuées sur le trafic Internet et la théorie des valeurs extrêmes. La section d'après se concentrera sur le trafic Internet et démontrera la pertinence de ce sujet au regard de l'époque dans laquelle nous vivons. La section 4, quant à elle, servira à poser les fondements et expliquer les concepts relatifs à la théorie des valeurs extrêmes utiles pour répondre à la question de recherche. Par la suite, après avoir brièvement décrit l'ensemble de données utilisé, nous allons entreprendre l'application de la méthode afin de répondre à la question de recherche. Enfin, la section 6 reprendra la conclusion, les limites et les recommandations.

2. Revue littéraire

Jusqu'à présent, le trafic Internet a fait l'objet d'un grand nombre de recherches et d'études. De nombreuses perspectives ont été explorées pour le trafic Internet telles que sa classification, sa mesure et sa structure. Certains articles scientifiques vont être passés en revue afin de fournir une vue d'ensemble de la littérature quant au trafic Internet ainsi qu'à la théorie des valeurs extrêmes dans ce contexte.

Tout d'abord, Tsourti et Panaretos (2004) se sont intéressés à l'analyse des valeurs extrêmes appliquée à des données relatives au télétrafic. Les auteurs utilisent la théorie des valeurs extrêmes telle que présentée dans la section 4. L'objectif était de découvrir le comportement des requêtes des utilisateurs par rapport au système. Selon les auteurs, l'ajustement de la capacité du système lui permettrait de fonctionner correctement et ainsi de pouvoir gérer les demandes les plus larges. Les deux méthodes utilisées ont confirmé une distribution à longues queues. Les auteurs en ont conclu que la présence de comportements à longues queues rendait invalide les hypothèses traditionnelles des modèles classiques, c'est-à-dire les modèles basés sur le théorème central limite. Cela confirme la potentielle grande utilité de l'analyse des valeurs extrêmes.

Ensuite, le trafic Internet a suscité l'attention d'Alasmar et al. (2019) quant à la répartition du volume du trafic Internet, puisque la distribution de la quantité du trafic par unité de temps avait peu été étudiée. Cette quantité peut pourtant s'avérer utile et importante dans la planification d'un réseau. Alasmar et al. (2019) ont dès lors cherché la distribution statistique correspondant le mieux au volume du trafic Internet, c'est-à-dire aux empreintes, au cours d'une période de temps déterminée. À cette fin, les auteurs se sont basés sur une quantité importante d'empreintes provenant d'une vaste étendue de temps. Trois distributions caractéristiques, à savoir normale, log-normale et Weibull, ont dès lors été testées sur diverses échelles de temps. Il a été conclu que, malgré la complémentarité de ces trois distributions, la distribution log-normale serait favorisée et non pas la distribution normale comme cela a longtemps été le cas. En effet, la loi log-normale avait une meilleure qualité de l'ajustement pour la plupart des empreintes analysées, pour une large variété de cas ainsi que pour toutes les échelles de temps considérées et étudiées.

Par ailleurs, l'analyse des données de trafic peut s'avérer être d'une grande aide pour prédire la qualité des télécommunications pouvant être dégradée à tout moment. Cela va avoir des implications dans le domaine du management. Uchida (2004) s'est attaqué à la problématique des graves dégradations de la qualité des télécommunications. Les données utilisées étaient des données de télétrafic récoltées sur un réseau réel. Une faible partie a été considérée comme des données connues pour l'analyse et les données restantes comme des données inconnues pour évaluer l'efficacité de la solution. De plus, les données connues ont été utilisées pour prédire des événements graves. La distribution généralisée de Pareto (comme présentée dans la section 4.b.iii.) s'est révélée estimer la distribution des queues des données connues beaucoup mieux que les distributions empirique ou log-normale.

L'auteur a ensuite voulu s'assurer de l'efficacité de cette distribution. Les résultats de l'analyse ont donc été employés pour examiner la valeur maximale du débit dans les données inconnues. Il en a résulté que la pratique utilisée offrait une meilleure estimation du débit maximal inconnu que celles employant la distribution empirique ou log-normale. De plus, les résultats se sont révélés être identiques même si d'autres données étaient utilisées. Selon Uchida (2004), prévoir des événements de dégradation graves, c'est-à-dire la détérioration de la qualité des télécommunications sur le réseau, avec une réduction du coût de mesure et de la quantité d'espace de stockage des données de trafic était donc possible.

Puis, la mesure du trafic Internet a attiré l'attention en raison de l'évolution d'Internet, dû au développement, à la croissance et à l'utilisation d'une plénitude d'applications de réseau. La compréhension du bon fonctionnement ou non d'un réseau local ou étendu constitue le but majeur de la mesure du trafic Internet. Williamson (2001) a démontré qu'au fil du temps, les chercheurs ont affiné les outils disponibles pour mesurer le trafic Internet. Selon l'auteur, un outil adapté permettrait de collecter des données et de déduire, par exemple, des informations à propos de la structure d'une application Internet ou du comportement d'un utilisateur Internet. Toutefois, le trafic Internet ne cesse d'évoluer et ce sur des intervalles de temps assez courts. Cette modification concerne la composition, le volume, les protocoles, les applications ainsi que les utilisateurs du trafic. Tel qu'expliqué par Williamson (2001), tout ensemble de données collectées à partir d'un réseau opérationnel ne représente qu'un instantané à un moment donné de l'évolution d'Internet. La mesure du trafic Internet est considérée par Williamson (2001) comme une méthodologie de recherche appliquée aux réseaux visant à comprendre le trafic de paquets[1] sur Internet. Un moyen de se mesurer aux recherches incessantes quant à la mesure et la compréhension du trafic Internet est d'arriver à déceler les éléments stables de sa structure.

Après, Färber et al. (1998) se sont aussi intéressés à la mesure du trafic Internet ainsi qu'à sa modélisation mais au niveau des réseaux d'accès. Selon les auteurs, en raison de l'inefficacité des méthodes habituelles pour dépeindre la charge de trafic actuelle, la conception et le dimensionnement corrects du réseau nécessiteraient le développement de nouveaux modèles de trafic. Une description facile du trafic devant être employée dans les modèles en vue de l'analyse et de la simulation des performances va être d'une grande aide pour les opérateurs de réseaux téléphoniques afin d'adapter leurs réseaux actuels à la hausse de la demande. Une raison de cette adaptation est l'influence des utilisateurs, accédant à Internet via le réseau téléphonique public commuté, et de leur comportement sur la performance des réseaux et services. Selon Färber et al. (1998), deux éléments semblaient être caractéristiques du trafic Internet, à savoir les longues durées d'attente ainsi que la grande variabilité du temps d'attente et du temps d'attente entre deux arrivées. Le système tarifaire téléphonique utilisé a largement impacté le comportement des utilisateurs. D'ailleurs, la modélisation a permis d'enregistrer l'attitude générale du trafic démontrant

[1] Selon le site Techno-Science, un paquet est une unité de transmission utilisée pour communiquer. (Source: Techno-Science.net, https://www.techno-science.net/definition/11437.html, consulté le 16/05/2021)

ainsi l'existence d'une variation des spécificités du trafic durant la journée. Dès lors, le dimensionnement du réseau se verrait faciliter avec un modèle présentant la charge du trafic uniquement durant les heures d'affluence du trafic Internet et téléphonique. Pour Färber et al. (1998), les résultats ne décrivaient pas nécessairement le trafic Internet général dû à l'utilisation de données empiriques d'un groupe spécifique d'utilisateurs.

En outre, la classification[2] du trafic Internet a fait l'objet de nombreuses études. Beaucoup d'auteurs ont désiré trouver LA méthode pouvant classer parfaitement le trafic Internet. Cependant, le degré de précision d'une telle classification n'est habituellement pas très élevé. De plus, cette classification représente un challenge en raison de l'existence d'un grand nombre et de l'émergence continue d'applications sur Internet. Moore et Zuev (2005) ont étudié cette classification et désiré prouver que la classification du trafic Internet par application pouvait atteindre une précision de plus de 95% avec les techniques appropriées. Le résultat de la méthode statistique de base utilisée, à savoir la méthode naïve bayésienne, ne fut pas très concluant en raison d'une précision estimée à environ 65%. Les auteurs ont néanmoins démontré que la précision globale de la classification augmentait lorsque ce classificateur était affiné. L'affinement s'est réalisé par l'utilisation de la théorie d'estimation de la densité de noyau et une méthode de pré-filtrage. La précision de la classification a dès lors atteint un niveau de 95% voire plus et était largement plus élevée comparée aux pratiques traditionnelles présentant un niveau de précision entre 50 et 70%. La stabilité temporelle de cette pratique a aussi été démontrée grâce à une meilleure réalisation de la classification du trafic pair-à-pair. La précision du modèle a donc augmenté au fil du temps.

Puis, Callado et al. (2009) se sont aussi penchés sur l'identification et la classification du trafic Internet et ont rassemblé certaines études abordant ces sujets. Ces derniers se sont d'abord focalisés sur la mesure du trafic Internet. Il en a résulté qu'il était parfois nécessaire d'employer des techniques d'échantillonnage tant certains liens pouvaient produire une quantité considérable de données. Callado et al. (2009) ont aussi abordé le sujet de l'analyse des flux et celui de la classification du trafic Internet. Ils en ont conclu l'absence d'une méthode assez précise mais surtout fiable pour effectuer la classification du trafic. Par conséquent, comme expliqué par Callado et al. (2009), aucune pratique n'a atteint, en même temps, une exactitude et une précision élevées dans la plupart des applications et ce malgré une plénitude de recherches.

Dainotti et al. (2012) ont continué ces recherches. Malgré l'amélioration de la précision et de l'efficacité des méthodes existantes, cette thématique est en constante évolution notamment à cause du changement perpétuel du comportement des applications Internet et de leur nombre. Selon les auteurs, la combinaison de toutes les méthodes existantes, un système de multi-classification, pourrait être la solution. Puisque chacune

[2] La classification du trafic Internet consiste à classer ensemble les applications Internet de même type. Le site CAIDA définit la classification du trafic Internet comme suit : nous utilisons l'expression classification du trafic pour décrire les méthodes de classification du trafic en fonction des caractéristiques observées passivement dans le trafic et en fonction d'objectifs de classification spécifiques. (Source: CAIDA, https://www.caida.org/archive/classification-overview/, Consulté le 16/05/2021)

d'entre elles performait de manière efficace sur un ensemble spécifique distinct de classes, une amélioration de la précision serait possible ainsi qu'une meilleure robustesse aux modifications de l'échantillon. La précision, la flexibilité et la vitesse se verraient donc augmenter grâce à cette multi-classification.

Ensuite, Alasmar et Zakhleniuk (2017) se sont intéressés à la qualité du service Internet. Les fournisseurs d'accès Internet emploient le dimensionnement des liens du trafic Internet afin d'approvisionner de manière appropriée la capacité de leurs liens de réseau. Ces derniers doivent être prêt à affronter les éventuelles coupures ou défaillances du réseau. La validation du réseau repose sur l'évaluation des mesures de la qualité de service dont les conditions sont définies dans un accord entre fournisseurs et utilisateurs, appelé accord sur le niveau de service. La planification de la bande passante du réseau est l'élément sur lequel les mesures de la qualité du service vont s'appuyer. La bande passante doit dès lors être suffisante pour l'approvisionnement. Les auteurs ont essayé d'établir le niveau souhaité de capacité de la bande passante le plus bas possible pour un lien du réseau et ont décidé d'employer l'approvisionnement en bande passante. Alasmar et Zakhleniuk (2017) ont démontré que la distribution du trafic Internet possédait des queues lourdes et ont ensuite cherché un modèle ajustant correctement le trafic. Les distributions à queues lourdes se sont avérées plus adaptées, optimales et précises pour décrire le trafic Internet. La distribution log-normale a assez bien ajusté le trafic capturé, alors que la répartition généralisée des valeurs extrêmes l'a extrêmement bien ajusté. La distribution normale s'est avérée ne pas du tout s'ajuster à la distribution du trafic capturé en raison d'une différence au niveau des queues. Il a été prouvé que l'approvisionnement en bande passante sur base des modèles log-normal et VEG[3] avait des résultats de performance très satisfaisants et acceptables et respectait dès lors les conditions de l'accord sur le niveau de service. Selon les auteurs, les approvisionnements en bande passante basés sur le modèle gaussien ne respectaient pas la capacité minimale souhaitée et violaient ainsi les conditions de l'accord.

Enfin, depuis peu, les scientifiques se sont attardés sur d'autres aspects de la théorie des valeurs extrêmes que ses distributions dans le but d'analyser le trafic Internet. Chen et al. (2016) se sont occupés de la modélisation et de la prédiction des taux de cyber-attaques extrêmes, à savoir le nombre d'attaques contre un système par unité de temps, apparaissant de manière plus fréquente sur Internet. Toutefois, les taux de cyber-attaques sont faibles la majeure partie du temps. L'allocation des ressources de défense doit dès lors se faire de manière dynamique et sur demande. Chen et al. (2016) ont proposé une nouvelle méthode pour la modélisation et la prédiction des taux de cyber-attaques extrêmes, le processus de points marqués se basant sur deux modèles. Il a été démontré que les modèles pouvaient prédire les taux extrêmes de cyber-attaques et étaient assez stables dans le temps. Chen et al. (2016) en ont conclu que l'approche des processus de points marqués pouvait prédire avec précision les taux extrêmes de cyber-attaques et possédait ainsi de bonnes performances d'ajustement et de prédiction.

[3] VEG signifie Valeur Extrême Généralisée.

3. Trafic internet

Au fur et à mesure du temps, Internet est devenu de plus en plus important. Ce dernier fait désormais partie intégrante de notre quotidien. Il serait impossible pour nous de nous en séparer tant il est riche en informations en tout genre. En raison de cet essor, le trafic Internet s'est considérablement intensifié, avec de plus en plus d'utilisateurs de jour en jour.

Le trafic Internet est défini, selon Wikipédia, comme la circulation des flux d'informations sur le réseau informatique mondial qu'est Internet, c'est-à-dire la circulation des données dans l'ensemble d'Internet ou dans certaines liaisons réseaux de ses réseaux constituants. De plus, il est précisé que ce trafic est alimenté par le trafic du Web, ainsi que par d'autres grands usages d'Internet tels que la messagerie électronique, les flux vidéo et audio et les réseaux pair-à-pair. De plus, le trafic d'un site web est défini, par le site Définitions Marketing, comme le nombre de visites ou visiteurs sur une période donnée.

Selon le site 'La Banque Mondiale' (**Annexe A.1**), en 2000, les utilisateurs d'Internet représentaient 6,526% de la population mondiale. En 2005, ils étaient 15,673%. En 2010, les utilisateurs d'Internet constituaient 28,706% de la population mondiale contre 41,546% en 2015. Enfin, en 2018, les utilisateurs d'Internet représentaient 50,763% (voir **Figure 1**). Cette évolution mondiale peut être visualisée par pays dans l'**Annexe A.2**.

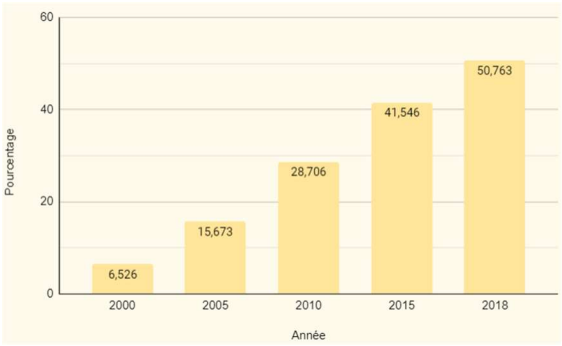


Figure 1: Évolution du pourcentage de la population mondiale en tant qu'utilisateurs d'Internet

De plus, le site 'Statista' nous fournit des informations à propos du 'temps passé par jour à utiliser Internet dans le monde en Janvier 2019, par pays' (**Annexe B**). Ce graphe nous montre la durée moyenne quotidienne pour 40 pays. D'après ce visuel, la population passe énormément de temps sur Internet. La moyenne mondiale est de 402 minutes, soit 6h42. Quant à la Belgique, ce temps est de 301 minutes, équivalentes à 5h01.

Ensuite, l'étude réalisée par WeAreSocial et Hootsuite permet de se rendre davantage compte de la popularité croissante d'Internet et des réseaux sociaux. En Janvier 2020, 4,54 milliards de personnes, soit 59% de la population mondiale, sont des utilisateurs d'Internet et 3,80 milliards de personnes, soit 49% de cette population mondiale, sont des utilisateurs actifs des réseaux sociaux (**Annexe C.1**). Le nombre d'utilisateurs d'Internet a en effet augmenté ces dernières années (**Annexe C.2**). En Janvier 2019, ce nombre s'élevait à 4,24 milliards de personnes et, en Janvier 2020, à 4,54 milliards de personnes, représentant ainsi

une augmentation de 7% par rapport à l'année précédente. De plus, le temps moyen passé sur Internet a de manière générale augmenté entre 2014 et 2019 pour atteindre une durée de 6h43 (**Annexe C.3**). Le temps quotidien passé sur Internet par pays nous démontre que ce temps quotidien moyen annoncé est largement dépassé par certains pays (**Annexe C.4**). Concernant les réseaux sociaux, le nombre de ses utilisateurs n'a fait que croitre depuis 2015 (**Annexe C.5**). Entre Janvier 2019 et Janvier 2020, une augmentation de 9,20% a été notée, passant de 3,48 à 3,80 milliards d'utilisateurs des réseaux sociaux. Le temps moyen passé quotidiennement sur ces derniers n'a fait qu'augmenter de manière progressive depuis 2014, atteignant une durée moyenne quotidienne de 2h24 en 2019 (**Annexe C.6**). Nous pouvons dès lors supposer que l'émergence des réseaux sociaux a indirectement favorisé la hausse de l'utilisation d'Internet et, par conséquent, la croissance du trafic Internet.

Cette étude peut être comparée à une autre étude à nouveau réalisée par WeAreSocial et Hootsuite pour le premier trimestre de 2020 (publiée en Avril). Tout d'abord, les proportions des utilisateurs d'Internet et des réseaux sociaux par rapport à la population mondiale sont les mêmes qu'au début de l'année (**Annexe D.1**). Cette différence réside dans la croissance de la population mondiale lors de ces quelques mois. Cependant, le nombre d'utilisateurs a augmenté. Il a été recensé 4,57 milliards d'utilisateurs d'Internet en Avril 2020. Pour les réseaux sociaux, nous sommes passés à 3,81 milliards d'utilisateurs actifs en Avril 2020.

De plus, entre Avril 2019 et Avril 2020, il y a eu une augmentation de 7,1% d'utilisateurs d'Internet et une hausse de 8,7% d'utilisateurs actifs des réseaux sociaux (**Annexe D.2**). La hausse des activités digitales entre Janvier et Avril 2020 est surtout due à la Covid 19. Certains utilisateurs d'Internet ont été questionnés (données collectées entre le 31 Mars 2020 et le 02 Avril 2020) à propos de leur temps passé sur différentes activités pendant la pandémie (**Annexe D.3**). 57% de cet échantillon a regardé plus de films et de séries via des services de streaming, 47% a passé plus de temps sur les réseaux sociaux, 46% a utilisé plus longtemps les services de messagerie et 35% a joué plus longtemps sur des jeux vidéo ou en ligne. Ensuite, l'étude s'est concentrée sur le changement de leurs habitudes en ce qui concerne les médias (**Annexe D.4**). Comparé au passé, 29% du panel a statué qu'il visionnait considérablement plus de films, séries et spectacles via les services de streaming. 23% de cet échantillon a passé beaucoup plus de temps à employer les réseaux sociaux et 24% a utilisé de manière plus significative les services de messagerie. Ces données nous confirment que la Covid 19 a augmenté le trafic Internet. Certaines personnes ont d'ailleurs affirmé maintenir dans le futur leurs nouvelles habitudes acquises lors de la pandémie (**Annexe D.5**). 20% des personnes interrogées ont déclaré s'attendre à visionner davantage de films et séries et 15% des personnes passant plus de temps sur les réseaux sociaux ont prévu de continuer de cette manière. Dès lors, même si le trafic Internet diminuera probablement après la pandémie, il aura néanmoins augmenté par rapport à l'année précédente étant donné la pérennité de ces nouvelles habitudes par certains. Les **Figures 2** et **3** fournissent un récapitulatif des éléments clés de l'utilisation d'Internet pour Janvier 2020 et Avril 2020, respectivement.

Figures 2 et 3: Vue d'ensemble de l'utilisation globale d'Internet en Janvier 2020 et Avril 2020, respectivement
(Source: BDM (Blog du Modérateur), https://www.blogdumoderateur.com/internet-reseaux-sociaux-2020/, consulté le 21 Janvier 2021, publié le 04 Février 2020 (gauche) et https://www.blogdumoderateur.com/internet-reseaux-sociaux-mobile-t1-2020/, consulté le 21 Janvier 2020, publié le 27 Avril 2020 (droite))

Cette enquête de WeAreSocial et Hootsuite confirme que la pandémie a provoqué une hausse majeure du trafic Internet. Lors du confinement, les vidéos en streaming, les vidéos conférence, le trafic des données mais aussi les appels Wi-Fi se sont largement accrus et davantage d'appareils étaient connectés au Wi-Fi. Cette situation inhabituelle nous permet de réaliser que le trafic Internet peut croître très rapidement et ne cesse de s'intensifier. Le trafic Internet peut aussi varier en fonction de certains événements, tels que de grands événements sportifs, le lancement d'une nouvelle série et une nouvelle saison tant attendue d'une série.

De plus, en Octobre 2018, Sandvine, société spécialiste des équipements de réseau, a publié un rapport indiquant que le streaming vidéo était l'application dominante en matière de création de trafic avec environ 58%. Dans cette catégorie, nous pouvons retrouver Netflix en première place avec un total de 26,6% du trafic Internet généré par le streaming vidéo alors qu'il ne reprend qu'environ 15% du trafic Internet global (**Annexe E.1**). En Octobre 2019, Sandvine a sorti un nouveau rapport sur le sujet, dans lequel il est précisé que 60,6% du trafic Internet (**Annexe E.2**) pour le premier semestre 2019 est consacré aux vidéos en streaming. D'ailleurs, ces vidéos en streaming et les jeux sont de plus en plus «tendance» à l'heure actuelle ainsi que le 4K et le 8K[4]. Par conséquent, la livraison d'un accès Internet à haut débit fait déjà l'objet de recherches pour pouvoir satisfaire au mieux les utilisateurs de plus en plus désireux. Pour satisfaire ce trafic, la rapidité d'exécution ne cesse d'évoluer. D'ailleurs, pour fournir un accès Internet à haut débit, l'industrie est en train de développer une plateforme '10G'. Cette plateforme livrera une vitesse de 10 Gigabits par seconde dans le monde pour satisfaire une demande de plus en plus importante en ce qui concerne la bande passante et une connectivité ininterrompue. Notre quotidien changera fortement tant dans l'apprentissage et le loisir que dans la vie et le travail par l'intégration du 10G, c'est-à-dire par cette redéfinition du rôle de la technologie.

[4] Le 4K et le 8K sont des formats d'image numérique ayant une définition, une résolution plus importante que le HD, respectivement, 4 fois et 8 fois plus importante que la résolution HD . Le 4K est aussi appelé le UHD, soit le Ultra Haute Définition. (Sources: CNET France, https://www.cnetfrance.fr/produits/sd-hd-ultra-hd-4k-8k-comprendre-les-definitions-destv-39786402.htm, consulté le 22 Janvier 2021)

Une question pourrait se lever 'Pourquoi le trafic Internet est-il pertinent dans le domaine du management?'. Le trafic Internet est un élément indispensable au webmarketing. En effet, le trafic Internet va nous permettre de mesurer l'efficacité d'une campagne marketing réalisée sur Internet. Comme mentionné précédemment, le trafic d'un site Internet représente le nombre de visiteurs du site Internet sur une période considérée. Il sera dès lors possible de déterminer le nombre de visiteurs étant venus sur le site lors de la durée de la campagne marketing et ainsi connaître son efficacité. D'ailleurs, Google Analytics est un outil très utile pour analyser le trafic. Certaines informations sur la quantité et la qualité du trafic peuvent être extraites, comme les pages avec le plus de trafic et le temps passé par un utilisateur sur une page du site. Ces données permettront à une entreprise de développer une stratégie particulière en vue de convertir de simples visiteurs en clients. De plus, si un site internet a un trafic élevé, les moteurs de recherche favoriseront davantage ces sites. En effet, ce facteur est considéré comme un signe de pertinence et de confiance par les moteurs de recherche. La quantité de trafic d'un site joue donc un rôle crucial dans le classement des sites par les moteurs de recherche, d'où l'importance d'un trafic fort et abondant pour une plus grande visibilité.

En outre, des données sur le trafic Internet vont permettre à une entreprise de gérer l'allocation de ses ressources quant à la gestion de son site Internet et/ou de ses systèmes. Avec les méthodes appropriées, une entreprise pourrait déterminer avec quel risque et/ou à quel moment son site Internet et/ou ses systèmes pourrai(en)t faire face à une panne. Sur base des résultats obtenus, une entreprise pourrait prendre des décisions quant au budget alloué à cette gestion et ainsi la capacité de son site et/ou de ses systèmes. En effet, une entreprise avec un budget élevé aura une capacité plus élevée alors que sa capacité sera plus faible si le budget alloué est plus faible. Par conséquent, grâce à une analyse du trafic Internet passé, une entreprise pourra réaliser des prédictions quant à la gestion de son site et/ou de ses systèmes et dès lors prendre les décisions les plus appropriées. C'est d'ailleurs ce point qui va nous intéresser dans ce travail.

4. Méthodologie

Cette partie théorique est développée sur base du livre de Coles (2001).

a. Rappel des concepts statistiques

Le cadre donné par Coles (2001) est le suivant. Les données $x_1, \dots, x_n$ sont des réalisations indépendantes d'une variable aléatoire X. Ces données représenteront le flux, à savoir les nombres de connexions mesurés à n instants. Sa fonction de densité appartient à une famille connue de distributions notée par $\mathcal{F} = \{f(x;\theta): \theta \in \Theta\}$. Le paramètre à estimer est θ et peut être soit un scalaire, soit un vecteur, avec un espace représenté par Θ. Dans notre cas, le paramètre θ sera un vecteur. La valeur réelle (mais inconnue) du paramètre θ_0 va pouvoir être approchée grâce à un estimateur, c'est-à-dire par une fonction de variables aléatoires. Nous appelons estimation la valeur prise par un estimateur pour un ensemble de

données étudié. Les valeurs fournies par les estimateurs pour les paramètres estimés doivent idéalement être proches de leur vraie valeur.

i. Estimation du maximum de vraisemblance

Le maximum de vraisemblance est une technique très utilisée pour estimer les paramètres d'un modèle. D'abord, chaque valeur de θ définit un modèle dans $\mathcal{F}$. Différentes probabilités sont ensuite attribuées par le modèle aux données observées. D'ailleurs, la fonction de vraisemblance représente la probabilité des données observées en fonction de θ. Quand des données se voient attribuer une probabilité élevée pour un certain θ, cela signifie que cette valeur de θ est vraisemblable. La méthode consiste dès lors à sélectionner le modèle avec la vraisemblance la plus élevée, c'est-à-dire le modèle attachant la probabilité la plus élevée aux données.

Si $x_1, \dots, x_n$ sont des réalisations indépendantes d'une variable aléatoire avec une fonction de densité $f(x; \theta_0)$, la fonction de vraisemblance est définie de la manière suivante:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

En conséquence, l'estimateur du maximum de vraisemblance $\hat{\theta}_0$ de θ_0 est défini comme la valeur de θ qui maximise la fonction de vraisemblance $L(\theta)$.

ii. Intervalles de confiance

L'écart-type d'un estimateur est une mesure implicite de sa précision. Au plus l'écart-type est petit, au plus la précision est élevée. La détermination d'un intervalle de confiance permet à la fois de donner un intervalle d'estimation et de s'assurer de la précision d'un estimateur. L'intervalle $\left[\hat{\theta}_l, \hat{\theta}_u\right]$ est un intervalle de confiance de niveau $(1-\alpha)$ pour θ_0 si

$$P\left\{\hat{\theta}_l < \theta_0 < \hat{\theta}_u\right\} = 1 - \alpha$$

Un intervalle de confiance va permettre de savoir avec quelle probabilité la vraie valeur du paramètre estimé se situe dans cet intervalle. L'étendue de l'intervalle de confiance est déterminée par le paramètre estimé, l'écart-type et le choix de α. D'ailleurs, le choix de α dépend entièrement d'une entreprise et repose sur un compromis entre largeur et niveau de confiance. Si un niveau de confiance élevé est désiré, α aura une petite valeur mais les intervalles générés seront larges. Si un niveau de confiance faible est voulu, α aura une grande valeur et les intervalles seront petits et plus étroits. Les valeurs α les plus populaires sont 0,05, 0,01 et 0,001, soit respectivement des niveaux de confiance de 95%, 99% et 99,9%.

b. Théorie des valeurs extrêmes

La théorie des valeurs extrêmes est appliquée dans de nombreux domaines, comme la finance et la météorologie. Les statistiques classiques se concentrent sur la moyenne alors que les statistiques extrêmes se focalisent sur l'étude des valeurs extrêmes des distributions.

La théorie des valeurs extrêmes, renommé «TVE» pour la suite de ce travail, a pour but de décrire les événements inhabituels et rares, tels que l'existence de valeurs extrêmes. Pour estimer ces valeurs, il est nécessaire de se baser sur les observations existantes en vue d'effectuer des déductions pour le futur, c'est-à-dire prédire à long terme à partir de données disponibles pour le court terme. Cela peut s'avérer dangereux puisque les distributions varient dans le temps. Ces déductions seront réalisables grâce aux modèles développés par la TVE. Cependant, ces modèles sont créés à partir d'un argument asymptotique et non pas sur une base empirique ou physique. Un modèle empirique se base sur des données existantes et n'est dès lors pas capable de réaliser des suppositions sur base de données extérieures à l'ensemble de données existant. Par contre, la TVE s'intéresse à des événements encore plus extrêmes que ce qui a déjà pu être observé dans le passé, à savoir des niveaux extrêmement élevés ou faibles. C'est la raison pour laquelle un argument asymptotique est utilisé à la place d'un modèle empirique. En résumé, la TVE va extraire des informations des valeurs extrêmes pour comprendre le comportement de ces dernières.

Cependant, comme toute méthode, la théorie des valeurs extrêmes possède certaines limites. L'utilisation d'un argument asymptotique est la première préoccupation. Les modèles obtenus ne doivent pas être considérés en tant que résultats exacts pour des échantillons finis. De plus, il faut faire attention à une potentielle non-stationnarité. L'argument asymptotique suppose que $X_1, \dots, X_n$ soient identiquement distribués mais, en réalité, ce n'est pas toujours le cas. Par exemple, les flux d'Internet sur une longue période ne suivront pas une même distribution car le flux augmente. Un risque de sous-estimation des niveaux de rendement et du risque associé est dès lors envisageable. Le second point concerne les circonstances sous lesquelles le modèle est développé. Ces dernières sont idéalisées ce qui s'avère ne pas être raisonnable pour une étude.

Ensuite, il est nécessaire de prendre en compte, à tout moment, quatre éléments essentiels quant à l'élaboration statistique des extrêmes.

1) Méthode d'estimation: une méthode d'estimation permet d'évaluer des paramètres inconnus en utilisant des données historiques. La méthode du maximum de vraisemblance est la plus adaptée dans l'étude des valeurs extrêmes car elle possède certaines propriétés d'inférence nécessaires dans notre cas, par exemple, elle s'adapte facilement aux modifications de modèles.

2) Quantification de l'incertitude: cela est très important puisque de nombreuses sources d'incertitude interviennent dans l'analyse des valeurs extrêmes. En effet, le moindre petit

changement du modèle peut conduire à de grandes variations causées par l'extrapolation. L'estimation de l'incertitude des niveaux extrêmes d'un processus peut être un paramètre de conception aussi important que l'estimation du niveau lui-même.

3) Diagnostic du modèle: le diagnostic du modèle est très important. L'extrapolation d'un modèle de valeurs extrêmes est uniquement rendue possible en raison de la base asymptotique à partir de laquelle il découle. La qualité de l'ajustement d'un modèle doit être évaluée parce que, si ce dernier ne reflète pas correctement les valeurs extrêmes, cela est très peu probable que l'extrapolation soit correcte.

4) Utilisation maximale de l'information: l'emploi de toutes les sources d'informations ainsi qu'une sélection appropriée du modèle et de l'inférence peut limiter les incertitudes.

i. Distribution des valeurs extrêmes

D'après Coles (2001), le modèle de base de la théorie des valeurs extrêmes se focalise sur le comportement statistique de M_n, le maximum d'une séquence de variables aléatoires indépendantes ayant une fonction de répartition F [5],

$$M_n = \max\{X_1, \dots, X_i, \dots, X_n\}$$

$X_1, \dots, X_n$ sont indépendants et identiquement distribués et représentent généralement les valeurs d'un processus mesuré sur une échelle de temps régulière afin que M_n représente le maximum du processus. De plus, en théorie, la distribution de M_n peut être dérivée pour toutes les valeurs de n,

$$P\{M_n \leq z\} = P\{X_1 \leq z, \dots, X_n \leq z\} = P\{X_1 \leq z\} \times \dots \times P\{X_n \leq z\} = \{F(z)\}^n = F^n(z)$$

Cependant, la distribution F de cette séquence est inconnue mais peut tout de même être estimée. La première étape est de chercher des familles de modèles appropriées pour F^n. Il est possible d'estimer cette dernière seulement avec les données extrêmes. Cela suit le même principe que pour l'approximation de la distribution des moyennes de l'échantillon par la distribution normale.

Pour n tendant vers l'infini, le comportement de F^n peut être étudié. De plus, F^n tend vers 0 quand n tend vers l'infini. Nous devons donc effectuer une renormalisation linéaire de M_n. Nous obtenons dès lors:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

$a_n > 0$ et b_n sont des constantes. F^n devient donc F^{n^*},

$$F^{n^*}(z) = P\left\{\frac{M_n - b_n}{a_n} \leq z\right\} = P\{M_n^* \leq z\}$$

[5] F représente la fonction de répartition, $F(x) = P\{X_i \leq x\}$.

Grâce à cette transformation, en plus des choix judicieux pour a_n et b_n, la recherche de distributions limites se tourne vers M_n^* à la place de M_n. D'ailleurs, plusieurs possibilités de distributions limites pour M_n^* existent.

Théorème 1 (Coles, 2001): s'il existe des séquences de constantes $a_n > 0$ et b_n de manière à ce que, quand n tend vers l'infini,

$$F^{n*}(z) \to G(z)$$

où G est une fonction de répartition non-dégénérative[6], alors G appartient à l'une des familles suivantes: Gumbel, Fréchet ou Weibull. Ces trois distributions sont reprises sous le nom de 'distributions de valeurs extrêmes'.

Par ce théorème, Coles (2001) affirme que les maxima de l'échantillon renormalisé M_n^* convergent en distribution vers une variable ayant une distribution dans l'une de ses familles. En d'autres termes, lorsque M_n peut être renormalisé avec a_n et b_n adéquats, la variable normalisée correspondante M_n^* a une distribution limite devant faire partie de l'une de ces trois familles de distributions de valeurs extrêmes. L'attribut saillant réside dans le fait que ces diverses familles sont les seules limites envisageables pour les distributions de M_n^*, peu importe la distribution F. De plus, les comportements respectifs de ces trois distributions limites diffèrent les uns des autres et représentent en réalité le comportent des valeurs extrêmes pour la fonction de distribution F de X_i. La distribution de Weibull a une limite supérieure finie. Les distributions de Fréchet et de Gumbel n'ont pas de limites supérieures et différent par rapport à leur densité. Pour la distribution de Fréchet et de Gumbel, respectivement, la densité de G décroît de manière exponentielle et de manière polynomiale. Par conséquent, le comportement des valeurs extrêmes est modélisé de manière largement distincte par les trois familles de distributions. Ces dernières peuvent être fusionnées en une seule famille de modèles, avec des fonctions de distribution:

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \qquad \text{(Équation 1)}$$

avec z tel que $1 + \xi\left(\frac{z-\mu}{\sigma}\right) \geq 0$.

Cela représente la famille des distributions des valeurs extrêmes généralisée, soit DVEG. μ désigne le paramètre de localisation, σ le paramètre d'échelle et ξ le paramètre de forme. Le paramètre ξ est le paramètre différenciateur entre ces trois familles de distributions. $\xi > 0$ renvoie à la famille Fréchet et $\xi < 0$ à la famille Weibull. $\xi = 0$ correspond à la famille Gumbel et est interprété comme la limite de $G(z)$ (Équation 1) quand ξ tend vers 0. Cette combinaison facilite l'exécution statistique. La déduction de ξ permet aux données elles-mêmes d'établir le type de comportement de queue le plus adapté. D'ailleurs, par souci de cohérence, il est nécessaire de reformuler le Théorème 1.

[6] La distribution G est dégénérative si elle est toujours soit 0, soit 1.

Théorème 2 (Coles, 2001): s'il existe des séquences de constantes $a_n > 0$ et b_n de manière à ce que, quand n tend vers l'infini,

$$F^{n*}(z) \to G(z)$$

pour une fonction de distribution non-dégénérative G, alors G est un membre de la famille des distributions des valeurs extrêmes généralisée, DVEG.

Dans le cas où n tend vers l'infini,

$$F^{n*}(z) \approx G(z)$$

Cette équation peut se transformer de manière équivalente en

$$F^{n}(z) \approx G^{*}(z)$$

où G^* est un autre membre de la famille des DVEG.

D'après Coles (2001), si le **Théorème 2** rend possible l'approximation de la distribution de M_n^* par un membre de la famille des DVEG pour un grand n, alors la distribution de M_n peut aussi être approchée par un membre distinct de cette même famille. Puisque les paramètres de la distribution doivent de toute façon être estimés, il n'est pas pertinent en pratique que les paramètres de la distribution G soient différents de ceux de G^*.

Malgré le large nombre soumis de méthodes, les techniques fondées sur la vraisemblance sont préférées pour l'estimation des paramètres. L'attractivité de cette pratique réside d'ailleurs dans son utilité polyvalente et son adaptabilité à la construction de modèles complexes. Cependant, l'emploi de la vraisemblance pourrait éventuellement poser problème au regard des conditions de régularité. Celles-ci sont réclamées en vue d'assurer la validité des propriétés asymptotiques habituelles attachées au maximum de vraisemblance. Étant donné que les extrémités de la DVEG sont des fonctions des paramètres, ces conditions ne sont pas respectées par le modèle VEG. Lorsque $\xi < 0$, une distribution possède une limite supérieure à $\mu - \frac{\sigma}{\xi}$ et n'a pas de limite inférieure. Lorsque $\xi > 0$, une distribution n'a pas de limite supérieure mais a une limite inférieure à $\mu - \frac{\sigma}{\xi}$. Par conséquent, les résultats asymptotiques types ne sont pas forcément prédicables en raison de la violation de ces conditions. Quand $\xi > -0{,}5$, les estimateurs de maximum de vraisemblance respectent les propriétés asymptotiques habituelles. Par contre, quand $\xi \leq -0{,}5$, les estimateurs de maximum de vraisemblance ne respectent pas ces propriétés et les distributions possèdent une queue supérieure bornée très courte. Néanmoins, il est relativement rare de faire face à ce cas de figure dans les applications de la théorie des valeurs extrêmes. En pratique, cette limite théorique ne constitue donc pas un problème.

La théorie des valeurs extrêmes a donc pour but de modéliser les événements rares, c'est-à-dire, dans le cas présent, des flux très élevés. Pour ce faire, nous allons considérer deux approches, le principe des maxima de blocs et le principe des dépassements de seuil.

ii. Principe des maxima de blocs

Le principe des maxima de blocs est une technique très utilisée en statistiques extrêmes. Son objectif est de sélectionner la valeur maximale sur une certaine période de temps fixe (un «bloc»), par exemple, sur une journée, sur un mois ou sur une année. Pour les maxima de blocs, le paramètre θ est un vecteur $\theta = (\mu, \sigma, \xi)$, avec $\Theta = \mathbb{R} \times (0, \infty) \times \mathbb{R}$.

La première étape est de sélectionner une taille de blocs appropriée pour l'ensemble de données utilisé. Pour une grande valeur de n, les données sont segmentées en blocs d'observations de longueur l. Les divers blocs ont donc une longueur identique. Un ensemble de maxima de blocs $M_{n,1}, \ldots, M_{n,m}$, m étant le nombre de blocs et $n \approx ml$, est dès lors obtenu et la DVEG peut être ajustée à cet ensemble. Cependant, une attention particulière doit être portée au choix de la taille des blocs. Ce choix est régi par un compromis entre biais et variance. Si les blocs sont trop petits, du biais sera présent dans l'extrapolation et l'estimation. Par contre, si les blocs sont trop grands, l'estimation fera face à une large variance en raison de la faible quantité de blocs. Souvent, la taille des blocs est sélectionnée de manière à ce que la période de temps considérée soit égale à un an. Les maxima deviendront des maxima annuels et l sera égal au nombre d'observations dans une année.

À partir de maintenant, $Z_1, \ldots, Z_m$ désignent les maxima de blocs, à savoir des variables indépendantes suivant une DVEG. Si les X_i sont indépendants, alors les Z_i sont également indépendants. Dès lors, les estimations des quantiles extrêmes de la distribution des maxima annuels sont obtenues en inversant l'équation 1,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] & pour\ \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\} & pour\ \xi = 0 \end{cases} \qquad \text{(Équation 2)}$$

où $G(z_p) = 1 - p$ et p est petit.

z_p est appelé le niveau (return level) associé à la période $\frac{1}{p}$. L'interprétation est que nous nous attendons à ce que le niveau z_p soit dépassé en moyenne une fois tous les $\frac{1}{p}$ ans. Autrement dit, cela signifie que z_p est dépassé par le maximum annuel pour une année donnée avec une probabilité p. Une inférence du niveau est obtenue en substituant dans l'équation 2 les estimateurs de maximum de vraisemblance des paramètres VEG, à savoir $\hat{\xi}$, $\hat{\sigma}$ et $\hat{\mu}$. Pour $0 < p < 1$, l'estimateur de z_p est

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p{}^{-\hat{\xi}}\right] & for\ \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma} \log y_p & for\ \hat{\xi} = 0 \end{cases}$$

où $y_p = -\log(1-p)$. Les faibles valeurs de p sont les plus désirables.

Cependant, l'interprétation des niveaux doit être réalisée de manière prudente, surtout quand cela concerne des longues périodes de rendement, c'est-à-dire avec un p très petit. L'incertitude de $\hat{z}_p$ sera caractérisée par un intervalle de confiance.

La dernière étape est le diagnostic du modèle. Cela consiste à vérifier la qualité de l'ajustement du modèle à nos données. Cette étape va permettre de déterminer si les conclusions réalisées à partir de nos données sont valables, étant donné que ces déductions dépendent de la pertinence du modèle estimé. La pertinence de ce modèle va dès lors être évaluée sur base de sa compatibilité avec les données. Cette qualité de l'ajustement d'un modèle VEG estimé est vérifiée grâce à l'utilisation de quatre graphiques.

D'abord, il y a les diagrammes de probabilités constituant une comparaison de la distribution empirique par rapport à la distribution ajustée. Il y a également les diagrammes de quantiles. Les mêmes informations sont fournies par ces deux graphes mais sont formulées sur des échelles distinctes. En revanche, la perception peut être tout à fait différente en fonction de l'échelle. Un ajustement peut paraître raisonnable sur l'une et pauvre sur l'autre. Les diagrammes de quantiles sont par contre capables de nous fournir de plus amples informations pour la région d'intérêt, à savoir pour de larges valeurs de z. Pour ces deux graphes, une défaillance du modèle VEG est affirmée si des écarts considérables sont observés. Les observations des diagrammes de probabilités et de quantiles devraient dès lors se situer près de la diagonale unitaire si G est une approximation raisonnable de F^n.

Le troisième graphique est celui du niveau (return level plot), consistant à représenter z_p en fonction de y_p, avec $y_p = -\log(1-p)$, sur une échelle logarithmique. En fonction de la valeur de ξ, le graphe prendra une forme différente. Si $\xi = 0$, le graphe sera linéaire. Si $\xi < 0$, le graphe sera convexe et, lorsque que p tendra vers 0, il possèdera une limite asymptotique à $\mu - \frac{\sigma}{\xi}$. Enfin, si $\xi > 0$, le graphe sera concave avec aucune limite finie. Des intervalles de confiance et les estimations empiriques des niveaux z_p sont également ajoutés afin de comparer les estimations empiriques aux estimations faites par le modèle. D'ailleurs, les diverses courbes doivent être en concordance pour confirmer l'ajustement du modèle VEG à nos données. Si cela n'est pas le cas, le modèle estimé est alors défaillant. Par conséquent, nous pouvons dire de ces trois graphiques qu'ils confrontent les estimations empiriques aux estimations du modèle. Enfin, le dernier graphique consiste à présenter l'histogramme des données avec la fonction de densité du modèle VEG estimé. Cependant, la grande variabilité d'un histogramme (en fonction des classes choisies) rend ce graphe moins intéressant, moins informatif, plus subjectif et plus difficile à interpréter. Les autres graphes sont donc généralement préférés.

De manière globale, toute analyse des valeurs extrêmes a une préoccupation majeure, la quantité limitée de données. Étant donné que les extrêmes sont des événements rares, les estimations qui en découlent ont une grande variabilité. Quant à la méthode des maxima de blocs, cette dernière ne prend en considération qu'un seul maximum pour un bloc.

Néanmoins, il est possible de trouver dans un bloc d'autres événements aussi rares et extrêmes que le maximum enregistré pour ce même bloc. Cela engendre une perte d'informations précieuses. C'est pourquoi, modéliser uniquement les maxima de bloc est une approche inefficace de l'analyse des valeurs extrêmes si d'autres données sur les extrêmes sont disponibles (Coles, 2001). Cette limite a conduit à deux autres caractérisations générales bien connues de la théorie des valeurs extrêmes. Nous allons nous intéresser à une seule d'entre elles, à savoir les dépassements d'un seuil élevé.

iii. Principe des dépassements de seuil

L'objectif de ce principe est de sélectionner toutes les valeurs dépassant un seuil important. Cette méthode va fixer un seuil et va considérer chaque observation dépassant ce seuil comme extrême. Cette approche s'oppose dans ce sens à l'approche des maxima de blocs. Pour ce deuxième principe, le paramètre θ est un vecteur $\theta = (\sigma, \xi)$, avec $\Theta = (0, \infty) \times \mathbb{R}$.

Pour rappel, la séquence de données $X_1, \dots, X_n$ est constituée de variables aléatoires indépendantes et identiquement distribuées. Ces variables possèdent une fonction de répartition F. Les observations X_i dépassant un seuil élevé u sont considérées comme extrêmes. En dénotant un terme de la séquence X_i par X, une description du comportement aléatoire des événements extrêmes est donnée par la probabilité conditionnelle suivante:

$$P\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0$$

À nouveau, la distribution parent F est inconnue, ainsi que celle des dépassements de seuil. Par conséquent, des recherches sont réalisées afin de déterminer des estimations de cette distribution valables pour des seuils élevés. Cela est équivalent au fait que la DVEG est employée comme approximation de la distribution des maxima.

Avant toute chose, il est important de définir la distribution propre à cette méthode, la distribution Pareto généralisée, DPG, dont la fonction de répartition est notée par $H(y)$.

Théorème 3 (Coles, 2001): Soit $X_1, X_2, \dots$ une séquence de variables aléatoires indépendantes avec une fonction de distribution commune F, et

$$M_n = \max\{X_1, \dots, X_n\}$$

Dénotez un terme arbitraire de la séquence X_i par X, et supposez que F satisfait le Théorème 1, de sorte que pour un grand n,

$$P\{M_n \leq z\} \approx G(z)$$

où

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

pour certains μ, $\sigma > 0$ et ξ. Alors, pour u suffisamment grand, la distribution de $(X - u)$, conditionnelle à $X > u$, est approximativement

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \qquad \text{(Équation 3)}$$

défini sur $\{y : y > 0\}$ et $\left\{\left(1 + \frac{\xi y}{\tilde{\sigma}}\right) > 0\right\}$, où

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \qquad \text{(Équation 4)}$$

$\tilde{\sigma}$ est le paramètre d'échelle de la DPG et σ est le paramètre d'échelle de la DVEG. La distribution $H(y)$ peut être considérée comme une distribution limite lorsque u tend vers l'infini. Le **Théorème 3** implique que, si les maxima des blocs ont une distribution approximative G, alors les dépassements de seuil ont une distribution approximative correspondante de la famille Pareto généralisée.

De plus, il existe une certaine dualité entre les familles VEG et Pareto généralisée. En effet, les paramètres de la DVEG vont déterminer ceux de la DPG correspondante. Ces deux distributions partagent le même paramètre de forme ξ. Dans l'équation 4, ξ reste invariable et les changements de μ et σ ne bouleversent pas le calcul de $\tilde{\sigma}$ étant donné que μ et σ s'autocompensent. Cependant, $\tilde{\sigma}$ est bien affecté par un changement de seuil.

Ensuite, le comportement de la DPG est défini en grande partie par le paramètre ξ, comme c'est le cas pour la DVEG. Si $\xi < 0$, la distribution n'a pas de limite inférieure mais a une limite supérieure à $u - \frac{\tilde{\sigma}}{\xi}$. Par contre, si $\xi > 0$, la distribution n'a aucune limite. Dans le cas où $\xi = 0$, la distribution n'a pas de limites mais doit être interprétée de manière asymptotique, c'est-à-dire la limite de l'équation 3 lorsque ξ tend vers 0,

$$H(y) = 1 - exp\left(-\frac{y}{\tilde{\sigma}}\right)$$

avec $y > 0$. Cette distribution correspond à une distribution exponentielle avec paramètre $\frac{1}{\tilde{\sigma}}$.

Sur base de l'équation 4, nous pouvons constater que la valeur de $\tilde{\sigma}$ dépend du seuil, excepté quand $\xi = 0$. D'ailleurs, dans le Théorème 3, il existe une distinction entre le paramètre d'échelle de la DPG, $\tilde{\sigma}$, et celui de la DVEG, σ. À partir de maintenant, cette distinction est abandonnée. Dès lors, σ désignera le paramètre d'échelle de manière générale, quelle que soit la famille considérée.

L'ensemble de données $x_1, \ldots, x_n$ est constitué de données indépendantes et identiquement distribuées. Un seuil u est établi de telle sorte que les événements sont considérés comme extrêmes, donc comme dépassements, quand $\{x_i : x_i > u\}$. Nous désignons ces dépassements par $x_{(1)}, \ldots, x_{(k)}$ et sont définis de la manière suivante: $y_j = x_{(j)} - u$, avec

$j = 1, \ldots, k$. D'ailleurs, la distribution de ces dépassements de seuil indépendants observés y_j peut être approchée par un membre de la famille Pareto généralisée.

Avant de passer à la suite, une attention particulière doit être accordée au choix du seuil. À nouveau, un compromis entre biais et variance est nécessaire. Si le seuil est trop bas, cela engendrera du biais et la base asymptotique du modèle ne sera peut-être plus respectée. Si le seuil est trop haut, cela entraînera une variance élevée en raison d'un nombre faible de dépassements. La norme est donc de fixer un seuil aussi bas que possible mais la condition est une approximation raisonnable du modèle. Dès lors, la pratique consiste à ajuster la DPG à un certain nombre de seuils potentiels et d'étudier la stabilité des estimations des paramètres en vue de choisir un seuil pour lequel les paramètres sont relativement stables. De manière plus formelle, selon le Théorème 3, les dépassements d'un seuil u devraient également être approchés par une DPG si celle-ci s'avère être un modèle raisonnable pour les dépassements d'un seuil u_0, avec u plus élevé que u_0. Dans ce cas, les deux distributions possèdent le même paramètre de forme ξ. La valeur de la DPG étant témoignée par σ_u pour un seuil $u > u_0$, l'équation 4 implique:

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$

De cette manière, le paramètre d'échelle varie avec u, excepté si $\xi = 0$. En reparamétrant, le paramètre d'échelle Pareto généralisé devient:

$$\sigma^* = \sigma_u - \xi u$$

Grâce à cette transformation, le paramètre reste constant par rapport à u. Cela suggère qu'au-dessus d'un seuil u_0, les estimations de σ^* et de ξ devraient rester stables. Cela sous-entend de représenter à la fois $\hat{\sigma}^*$ et $\hat{\xi}$ en fonction de u et, ensuite, de choisir un seuil u_0, perçu comme la plus petite valeur de u pour laquelle les estimations possèdent une certaine stabilité. La méthode du maximum de vraisemblance est utilisée pour obtenir ces estimations $\hat{\sigma}^*$ et $\hat{\xi}$.

Nous faisons ensuite l'hypothèse que la DPG est un modèle convenable pour les dépassements d'un seuil u par la variable X. Pour $x > u$,

$$P\{X > x | X > u\} = \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

Après arrangement, nous obtenons

$$P\{X > x\} = \zeta_u \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \qquad \text{(Équation 5)}$$

où $\zeta_u = P\{X > u\}$. ζ_u est la probabilité qu'une observation individuelle dépasse le seuil u. Par conséquent, le niveau x_m (return level) étant dépassé en moyenne une fois toutes les m observations est la solution de

$$\zeta_u \left[1 + \xi \left(\frac{x_m - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{m}$$

En réarrangeant,

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] & si\ \xi \neq 0 \\ u + \sigma \log(m\zeta_u) & si\ \xi = 0 \end{cases}$$

à condition que m soit suffisamment grand pour assurer que $x_m > u$.

De plus, de la même manière que pour les maxima de bloc, sur le graphe entre x_m et m, une linéarité sera présente si $\xi = 0$. Si $\xi > 0$, le graphe sera plutôt concave alors qu'il sera plutôt convexe si $\xi < 0$.

Dans l'approche des maxima annuels, le niveau sur N années sera dépassé en moyenne une fois tous les N ans. Si n_y est le nombre d'observations pour une année, dès lors $m = N \times n_y$. Le niveau sur N années est

$$z_N = \begin{cases} u + \frac{\sigma}{\xi}\left[\left(Nn_y\zeta_u\right)^{\xi} - 1\right] & si\ \xi \neq 0 \\ u + \sigma log\left(Nn_y\zeta_u\right) & si\ \xi = 0 \end{cases}$$

Par après, le niveau z_N peut être estimé en remplaçant les paramètres σ et ξ par leurs estimateurs de maximum de vraisemblance. Par contre, l'estimateur de ζ_u peut être calculé empiriquement:

$$\hat{\zeta}_u = \frac{k}{n}$$

où k est le nombre d'observations dépassant u. Puisque le nombre de dépassements de u suit la distribution binomiale Bin(n, ζ_u), $\hat{\zeta}_u$ représente aussi l'estimation du maximum de vraisemblance de ζ_u.

Comme pour la méthode des maxima de blocs, la dernière étape est de vérifier la qualité de l'ajustement du modèle Pareto généralisé estimé en se basant sur les quatre graphiques dédiés à ce sujet. Pour les diagrammes de probabilités et de quantiles, les points devraient à nouveau suivre la diagonale unitaire et une déviation prouverait une défaillance du modèle.

5. Application de la méthode des valeurs extrêmes

Nous allons d'abord passer en revue le jeu de données et, ensuite, nous nous focaliserons sur l'analyse et les résultats de l'application (de notre méthode).

a. Jeu de données

Le jeu de données provient du site Internet Kaggle[7] et se prénomme «Consumer traffic network». Ce dernier rassemble des données provenant de dix adresses IP locales recueillies sur une période de trois mois, soit du 01 Juillet 2006 au 30 Septembre 2006. Il faut noter que «la moitié de ces adresses ont été, à un moment donné, compromises au cours de cette durée et sont devenues membres de divers botnets[8]» (Kaggle). Une adresse IP compromise signifie qu'un hacker a accès à la machine et possède un contrôle sur cette dernière. De plus, le jeu de données possèdent 20 803 observations. Ainsi, une ligne de l'ensemble de données représente une observation.

L'ensemble de données est composé de quatre variables. La première variable représente la date à laquelle l'observation est enregistrée. Elle se trouve sous la forme année-mois-jour. Ensuite, nous avons l'adresse IP sur laquelle l'observation est apparue, répertoriée comme un nombre entier. Chaque nombre entier, de 0 à 9, est associé à une et une seule adresse IP. La troisième variable signale le numéro du registre autonome à distance. Ce numéro est en réalité un entier spécifiant le fournisseur d'accès Internet distant. Enfin, la dernière variable fait référence au flux lié à l'observation enregistrée, c'est-à-dire le nombre de connexions.

Enfin, il est possible d'interpréter d'une certaine manière chaque ligne du tableau. Pour un jour donné, un fournisseur d'accès Internet spécifique se connecte à une adresse IP qui va enregistrer son activité, c'est-à-dire son nombre de connexions pour ce jour précis.

b. Analyse

En vue de répondre à notre question de recherche, la variable d'intérêt est la variable 'flux'. Cette dernière décrit le nombre de connexions relatif à un fournisseur d'accès Internet particulier pour un jour donné. Plusieurs observations peuvent donc apparaître sur la durée d'une même journée. De plus, avant d'entrer dans l'application des différentes méthodes discutées ci-dessus, nous allons d'abord décrire l'ensemble de données de manière globale. Cela permettra d'avoir une vue d'ensemble des données en notre possession.

[7] L'ensemble de données trouvé sur Kaggle provient de http://statweb.stanford.edu/~sabatti/data.html.

[8] D'après le site Enisa, un botnet est défini comme un ensemble d'ordinateurs infectés par des bots. Un bot est un logiciel malveillant qui reçoit des ordres d'un maître. (Source: ENISA, https://www.enisa.europa.eu/topics/csirts-in-europe/glossary/botnets, Consulté le 07 Mai 2021)

i. Statistiques descriptives

Tout d'abord, une manière de commencer l'analyse des données est de connaitre ses statistiques descriptives. Le tableau suivant regroupe certaines de ces statistiques.

	Moyenne	Variance	Écart-type	Minimum	Quartile 1	Médiane	Quartile 3	Maximum
Flux	93,91	33 235 170	5 764,995	1	1	2	8	784 234

Tableau: Statistiques descriptives de la variable flux[9]

Ces diverses statistiques du flux nous permettent de remarquer que les données varient énormément. D'après ce tableau, nous pouvons constater que la valeur minimale est de 1, alors que la valeur maximale est de 784 234. Nous pourrions nous dire que les valeurs du flux sont bien réparties entre ces deux valeurs mais ce n'est point le cas. Une explication peut être fournie par les quartiles. D'une part, le quartile 1 a une valeur de 1. Cela signifie qu'un quart de nos données ont une valeur de 1, ce qui représente un ratio important. D'autre part, le quartile 3 a une valeur de 8. Cela indique que trois quarts des données ont une valeur inférieure ou égale à 8. En prenant en compte ces deux valeurs comparées à la valeur maximale, cela démontre que la plupart de nos données ont de très petites valeurs et se concentrent ensemble. Cet aspect peut être mis en lumière par la médiane. Cette dernière a une valeur de 2, révélant dès lors que 50% de nos données ont une valeur inférieure ou égale à 2.

De plus, l'écart-type des données nous permet de connaitre leur dispersion. La valeur de cet écart-type vaut ici 5 764,995. D'ailleurs, la moyenne vaut ici 93,910 alors que la médiane vaut 2. Cette grande différence montre que la moyenne n'est pas représentative de nos données. Cela peut s'expliquer par le fait que les grandes valeurs présentes dans les données vont avoir un poids important dans le calcul de la moyenne et donc la fausser.

Cependant, à partir de ces statistiques, nous pouvons déduire la présence d'une forte asymétrie dans nos données. Pour confirmer cette hypothèse, nous pouvons calculer un coefficient d'asymétrie γ, égale à 126,131. Celle-ci nous révèle donc une asymétrie de nos données, c'est-à-dire que les extrémités sont différentes les unes des autres. Les données possèdent plus particulièrement une asymétrie à droite étant donné la valeur positive. Ensuite, un calcul du kurtosis nous sera très utile. En effet, le kurtosis permet de mesurer la longueur (lourdeur) des queues (tail heavyness). La valeur obtenue pour le kurtosis est de 16 696,890. Cette dernière est largement supérieure à la valeur 3, cela signifie que la distribution est très pointue et accompagnée de queues de distributions très épaisses, c'est-à-dire des extrémités très longues. En d'autres termes, cela révèle que la distribution

[9] Les quantiles divisent un jeu de données en intervalles contenant le même nombre de données. Les quartiles divisent un jeu de données en 4 parts égales. Le quartile 1 correspond donc à la valeur pour laquelle 25% des données ont une valeur inférieur ou égale à cette valeur alors que le quantile 3 correspond à la valeur pour laquelle 75% des données ont une valeur inférieur ou égale à cette valeur. La médiane, ou quartile 2, sépare les données en 2 parts égales, 50% des données avec une valeur inférieure et 50% avec une valeur supérieure.

possède de très grandes valeurs, soit des valeurs extrêmes. De plus, cela nous suggère que nous allons très probablement trouver un ξ positif. Un histogramme du flux (**Figure 4**) nous fournit une visualisation de cette asymétrie à droite. La ligne rouge sur cet histogramme représente la densité du logarithme du flux. Cette représentation est un estimateur par noyaux de la densité. Pour une meilleure visualisation, nous avons pris le logarithme des données et de la densité. Cependant, il est nécessaire d'attirer l'attention sur un élément important. De manière générale, le logarithme est utilisé en vue de stabiliser, c'est-à-dire diminuer, une asymétrie. Mais malgré ce logarithme, nous voyons toujours une forte asymétrie et un fort kurtosis.

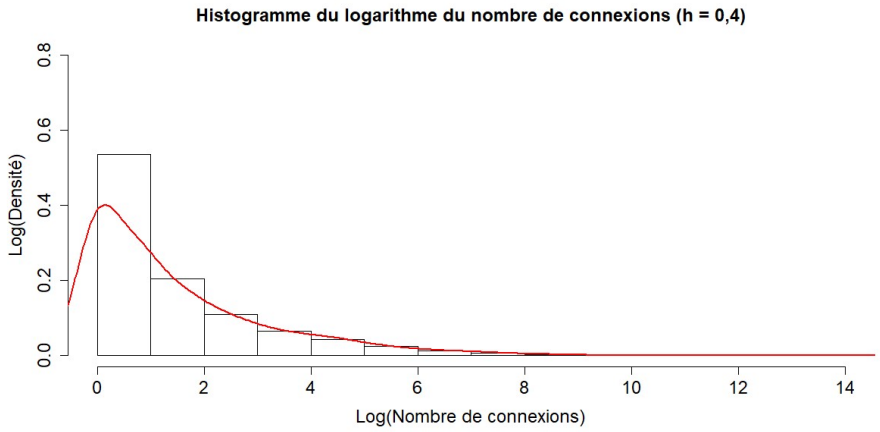


Figure 4: Histogramme du logarithme de la variable flux avec la densité du logarithme du flux représentée en rouge

Nous pouvons nous pencher désormais sur les relations entre le flux et les autres variables. Nous allons prendre, pour certains graphes, une échelle logarithmique pour l'axe des ordonnées afin d'obtenir des graphes lisibles. L'avantage majeur de cette échelle logarithmique réside dans le fait que les données gardent leur valeur d'origine.

D'abord, le nuage de points entre la variable flux et la variable adresse IP (**Figure 5**) nous permet d'avoir une vue d'ensemble du flux pour chaque adresse IP de notre ensemble de données. Nous pouvons constater que, pour certaines adresses, le flux est énormément plus important, démontrant ainsi la présence de ce que nous appelons des valeurs extrêmes.

Toutefois, si nous mettons en relation la variable registre et la variable adresse IP, cela va nous montrer que les fournisseurs d'accès à Internet sont répartis de manière équitable entre les différentes adresses IP. Ensuite, le graphe entre la variable flux et la variable registre va simplement nous donner le flux pour chaque fournisseur d'accès à Internet. En outre, la relation de la variable date avec la variable registre ainsi que celle avec la variable adresse IP n'ont pas de sens dans le cas présent.

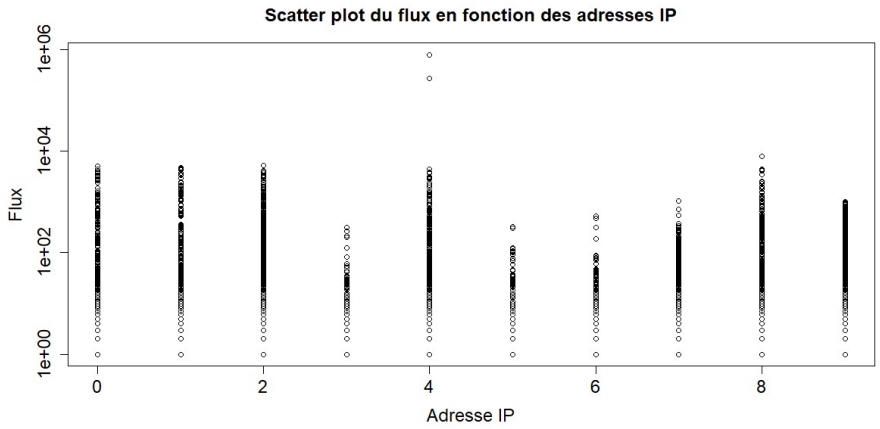


Figure 5: Nuage de points entre la variable flux et la variable adresse IP avec l'axe des ordonnées en échelle logarithmique

En revanche, la corrélation entre la variable date et la variable flux (**Figure 6**) nous est très utile. Sur cette figure, nous pouvons observer le flux pour chaque jour de la période considérée. Pour certains jours, il y a un nombre important de connexions. Pour d'autres, le niveau de connexions n'est pas trop élevé. Ce graphe permet de confirmer l'indépendance de nos données. En effet, ces dernières ne forment pas de clusters.

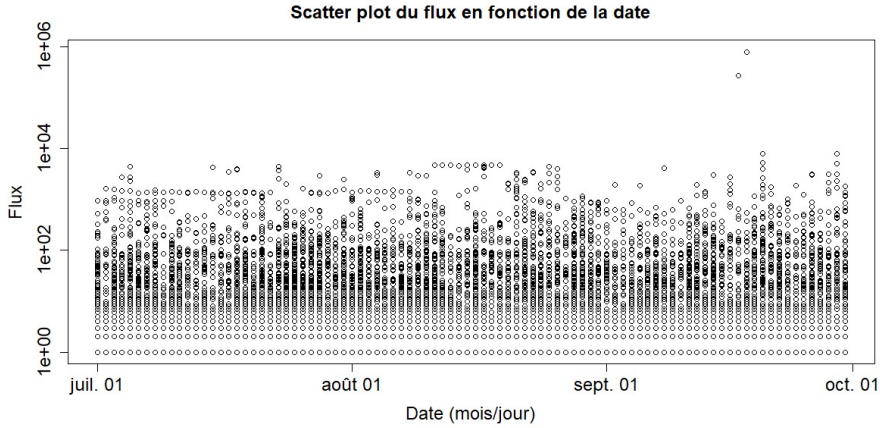


Figure 6: Représentation de la variable flux en fonction de la variable date, classifié par jour, avec l'axe des ordonnées en échelle logarithmique

Étant donné le nombre important d'observations, pour la suite de cette analyse, nous allons nous focaliser sur deux adresses IP, l'adresse IP 2 avec davantage de flux et l'adresse IP 3 avec un flux moins important. Par conséquent, le flux relié à ces adresses IP représentera nos données, $X_1, \dots, X_{n_1}$ pour l'adresse IP 2 avec $n_1 = 2\,416$ et $Y_1, \dots, Y_{n_2}$ avec $n_2 = 1\,186$. Le nombre de connexions moyen par jour de l'adresse IP 2 sera de 26,261 et celui de l'adresse IP 3 est de 13,326. De plus, prendre deux adresses IP avec des niveaux de flux différents représente un avantage certain. Cela permettra de savoir si les pratiques employées seront valides peu importe la quantité de flux impliquée. De plus, il est instructif de mentionner que l'adresse IP 2 n'a pas été compromise tandis que l'adresse IP 3 a été compromise. Un flux plus élevé pour une adresse IP ne veut pas forcément dire qu'une activité anormale est présente. Pour preuve, l'adresse IP 3, ayant été compromise, a un flux

bien moindre que le flux de l'adresse IP 2. L'aspect anormal est plutôt par rapport à l'activité moyenne d'une adresse IP. Il n'est donc pas vraiment possible de voir si une adresse IP a été compromise, ou non, aussi facilement. C'est justement cela la difficulté. La **Figure 7** et la **Figure 8** représentent, respectivement, le flux pour l'adresse IP 2 et le flux pour l'adresse IP3. D'une part, le flux de l'adresse IP 2 peut s'avérer très élevé de temps à autre, avec une valeur maximale de 5 214. D'autre part, contrairement à l'adresse IP 2, le flux de l'adresse IP 3 est relativement constant et faible, avec une valeur maximale de 313, soit bien moindre.

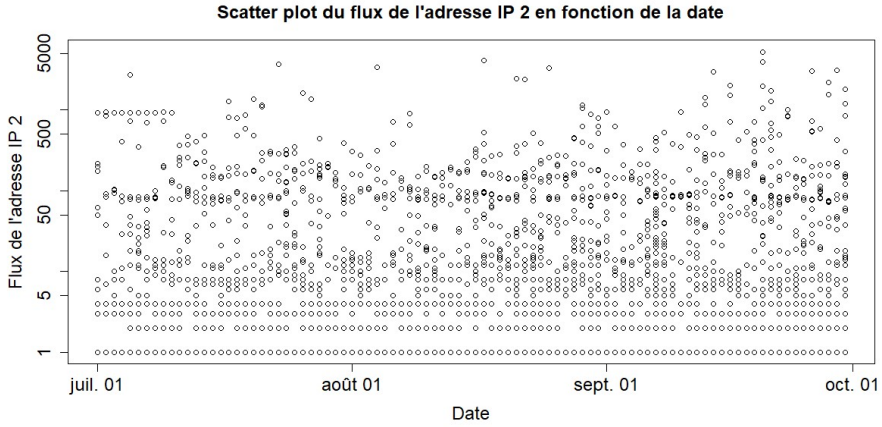


Figure 7: Représentation de la variable flux pour l'adresse IP 2 en fonction de la variable date, classifié par jour, avec l'axe des ordonnées en échelle logarithmique

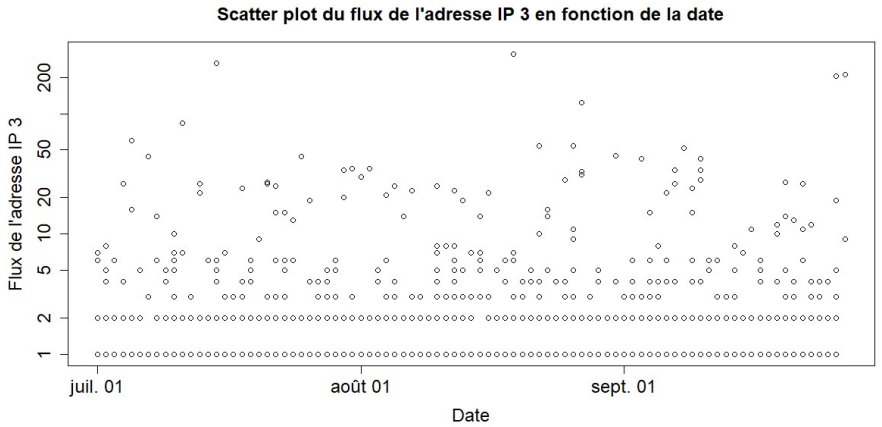


Figure 8: Représentation de la variable flux pour l'adresse IP 3 en fonction de la variable date, classifié par jour, avec l'axe des ordonnées en échelle logarithmique

Comme mentionné précédemment, le calcul de γ va nous permettre de déterminer si une asymétrie est présente dans nos données. Les valeurs obtenues pour les adresses IP 2 et 3 sont, respectivement, de 8,899 et de 13,415. Ces résultats étant positifs, les distributions de ces deux adresses IP sont dès lors asymétriques à droite. Ensuite, le kurtosis permet de déterminer la mesure de la longueur des queues, soit indirectement la présence de valeurs extrêmes dans les données. Le kurtosis de l'adresse IP 2 vaut 101,665 alors que celui de l'adresse IP 3 est de 211,004. Cela veut dire que les distributions des deux adresses IP sont pointues avec des extrémités longues. Par conséquent, nous pouvons en conclure l'existence de valeurs extrêmes. La **Figure 9** représente l'asymétrie pour l'adresse IP 2 et la **Figure 10**

l'asymétrie pour l'adresse IP 3. Ces figures représentent les histogrammes du logarithme du flux pour l'adresse IP correspondante avec leur densité de distribution en rouge. Une fois encore, malgré le logarithme supposé diminuer l'asymétrie, une forte asymétrie et un fort kurtosis sont encore présents.

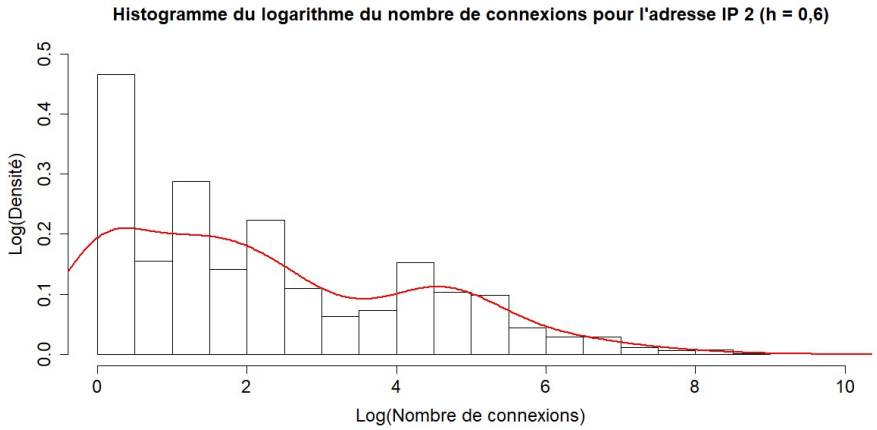


Figure 9: Histogramme du logarithme de la variable flux pour l'adresse IP 2 avec la densité du logarithme du flux de l'adresse IP 2 représentée en rouge

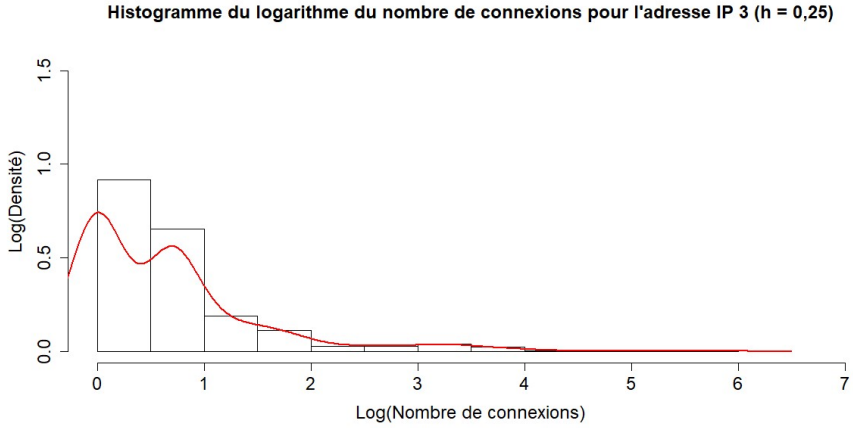


Figure 10: Histogramme du logarithme de la variable flux pour l'adresse IP 3 avec la densité du logarithme du flux de l'adresse IP 3 représentée en rouge

Les interprétations de ces deux adresses IP nous démontrent que les tendances observées pour le flux global se maintiennent pour les deux adresses IP sélectionnées. Nous pouvons donc nous permettre de nous concentrer uniquement sur ces deux adresses IP.

Enfin, pour conclure cette première partie de l'analyse, nous pouvons affirmer que les données nous suggèrent que $\xi > 0$. Comme vu dans la partie théorique (section 4.b.i.), ceci signifie que la distribution n'a pas de limite supérieure et peut prendre n'importe quelle valeur.

ii. Théorie des valeurs extrêmes

Comme mentionné précédemment, il existe diverses familles de distributions, comme la distribution Weibull, normale, exponentielle et log-normale. Il est dès lors pertinent d'appliquer ces distributions à nos données.

Les **Figures 11a**, **11b**, **11c** et **11d** représentent les histogrammes du flux de l'adresse IP 2, respectivement, avec une représentation de la distribution Weibull, exponentielle, normale et log-normale. De plus, les graphes ont été coupés pour avoir une meilleure visibilité au niveau des non-extrêmes afin de déterminer de l'ajustement potentiel de ces distributions. Il est relativement difficile de déduire quoi que ce soit à partir de ces graphes. Néanmoins, nous pouvons déjà constater que la distribution normale ne correspond en aucune manière à nos données.

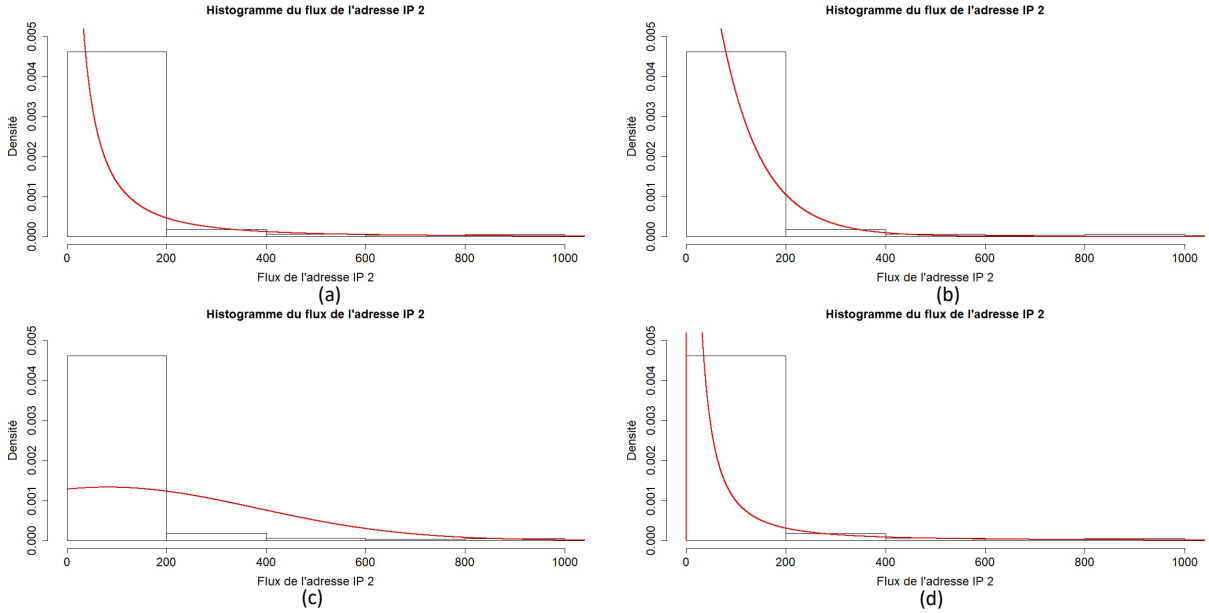


Figure 11: Histogrammes de la variable flux pour l'adresse IP 2 avec représentation de (a) la distribution Weibull, (b) la distribution exponentielle, (c) la distribution normale et (d) la distribution log-normale.

Les diagrammes quantiles-quantiles, ou QQ plots, sont un procédé plus adéquat pour déterminer ou plutôt évaluer la qualité de l'ajustement d'une distribution à un modèle théorique ou une autre distribution. Ces diagrammes vont représenter les quantiles de nos données (les points), axe des ordonnées, en fonction des quantiles de la distribution d'intérêt (la droite), axe des abscisses. Grâce à cette comparaison, cet outil permettra une meilleure visibilité, une meilleure déduction et surtout l'obtention de résultats plus précis. Les **Figures 12a**, **12b**, **12c** et **12d** représentent les QQ plots avec la distribution Weibull, exponentielle, normale et log-normale, respectivement. Ces graphes nous permettent de constater que les distributions Weibull et exponentielle s'ajustent plutôt bien aux données, à l'exception des extrêmes. De manière globale, la plupart de nos données sont correctement ajustées, hormis les extrêmes. En d'autres termes, ces distributions se révèlent finalement ne pas être adaptées à notre but. En effet, nous cherchons à modéliser les extrêmes, c'est-à-

dire les flux anormalement grands. C'est pourquoi, les distributions Weibull et exponentielle ne nous sont pas utiles. Celles-ci pourront seulement modéliser les flux moyens. Ensuite, en ce qui concerne la distribution normale, notre supposition relevée précédemment est confirmée. En effet, l'ajustement n'est pas du tout satisfaisant. Cette même conclusion peut être étendue à la distribution log-normale. En effet, cette dernière ne s'ajuste pas correctement à nos données. Cela vient d'ailleurs confirmer les dires de certains articles scientifiques démontrant que les données avec des extrêmes n'ont pas une distribution normale ou log-normale. En somme, la distribution de nos données possède en réalité davantage d'asymétrie que ces quatre familles de distributions.

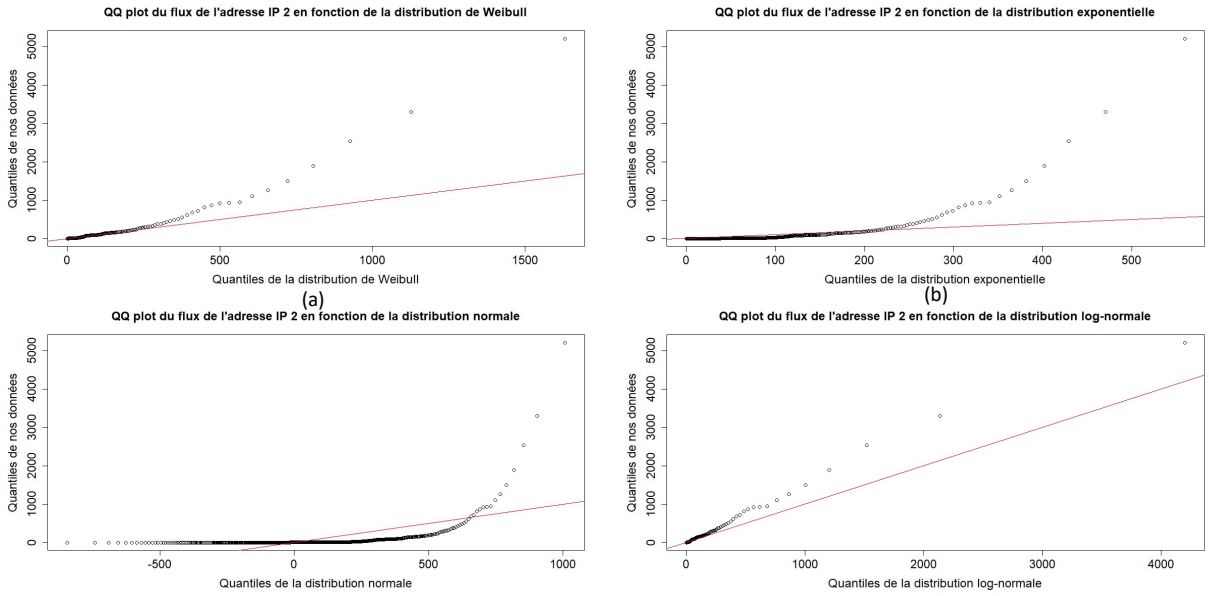


Figure 12: Diagrammes quantiles-quantiles du flux de l'adresse IP 2 par rapport à (a) la distribution Weibull, (b) la distribution exponentielle, (c) la distribution normale et (d) la distribution log-normale.

En outre, la même analyse peut être réalisée pour le flux de l'adresse IP 3. Cependant, nous allons directement nous focaliser sur les diagrammes quantiles-quantiles étant donné qu'ils offrent une meilleure précision en vue d'estimer la qualité de l'ajustement d'une distribution à une autre distribution. Les **Figures 13a**, **13b**, **13c** et **13d** représentent les QQ plots, respectivement, avec la distribution Weibull, exponentielle, normale et log-normale. Tout comme pour l'adresse IP 2, les distributions Weibull et exponentielle s'ajustent assez correctement aux données, à l'exclusion des extrêmes. Les distributions normale et log-normale ne s'adaptent à nouveau pas à nos données. Par conséquent, des conclusions identiques peuvent être établies. Encore une fois, ces quatre familles de distributions ne correspondent pas à nos données. La distribution actuelle de nos données dispose de queues de distribution bien plus longues que celles de ces familles de distributions. Autrement dit, l'asymétrie est davantage importante pour la distribution de nos données.

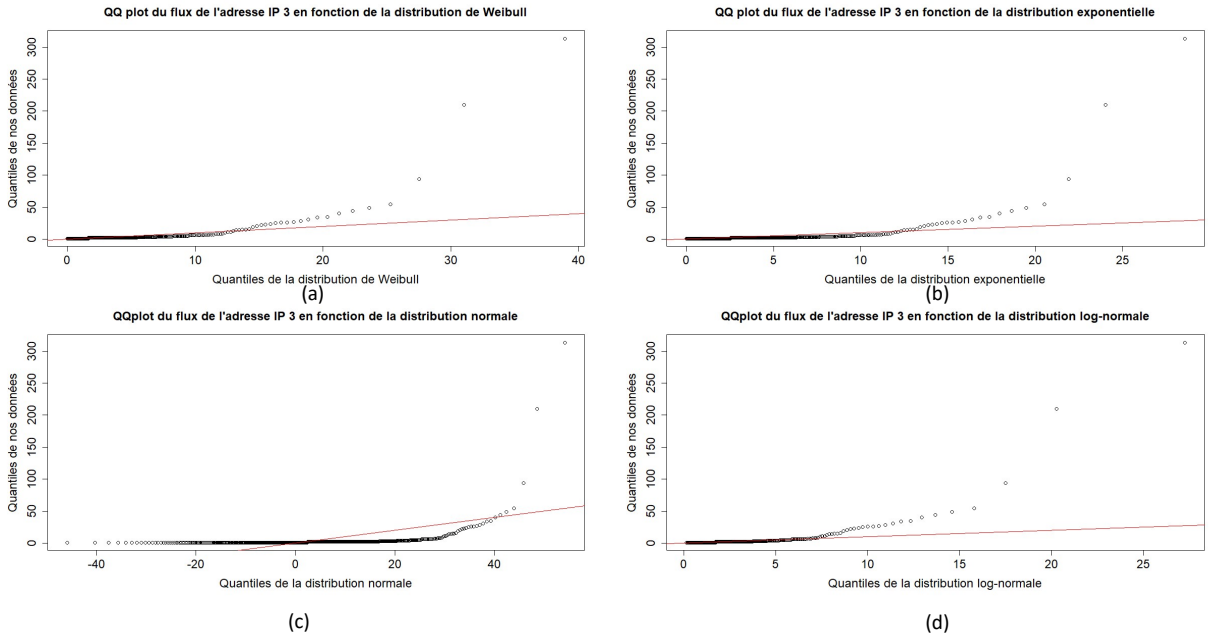


Figure 13: Diagrammes quantiles-quantiles du flux de l'adresse IP 3 par rapport à (a) la distribution Weibull, (b) la distribution exponentielle, (c) la distribution normale et (d) la distribution log-normale.

1. Maxima de blocs

a. Adresse IP 2

Tout d'abord, la première étape est de fixer une taille de blocs. Nous pourrions prendre des blocs d'une taille équivalente à un jour, une semaine ou même un mois. Considérer le flux par jour semble néanmoins être un bon compromis entre quantité trop faible et quantité trop élevée de données. Cependant, en considérant un jour comme un bloc, la constance entre les blocs, c'est-à-dire une longueur identique, s'en voit peut-être compromise. En effet, le flux est variable d'un jour à l'autre. Le nombre de connexions peut être totalement différent. Mais cela ne pose pas de problèmes si nous y faisons attention dans nos interprétations. Dans notre situation, cette configuration par jour est la meilleure solution. En effet, si nous prenons des blocs de même taille, l'interprétation se verrait compliquée. En termes de temps, les blocs auraient eu des tailles différentes et la prévision d'un dysfonctionnement d'un système ou d'un site serait dès lors compliquée. C'est pourquoi, le flux par jour est donc notre mesure pour la longueur des blocs.

Ainsi, notre ensemble de données pour l'adresse IP 2 sera composé de 92 maxima de blocs. Cela signifie que la valeur maximale de chaque jour de la période considérée est gardée. Cette sélection s'effectue à partir du flux relatif à l'adresse IP 2. Ainsi, l'ensemble de maxima est créé et constitue à présent l'ensemble de données sur lequel nous allons travailler. La **Figure 14** montre une représentation de ces maxima de blocs. Il est possible de noter que les maxima sont assez bien répartis mais différent largement en termes de valeur en fonction du jour considéré. Par la suite, nous pouvons représenter ces maxima de blocs sur le graphe du flux global de l'adresse IP 2 (**Figure 15**). Cela donne une meilleure représentation de la répartition et la situation de ces maxima par rapport au flux général de l'adresse IP 2.

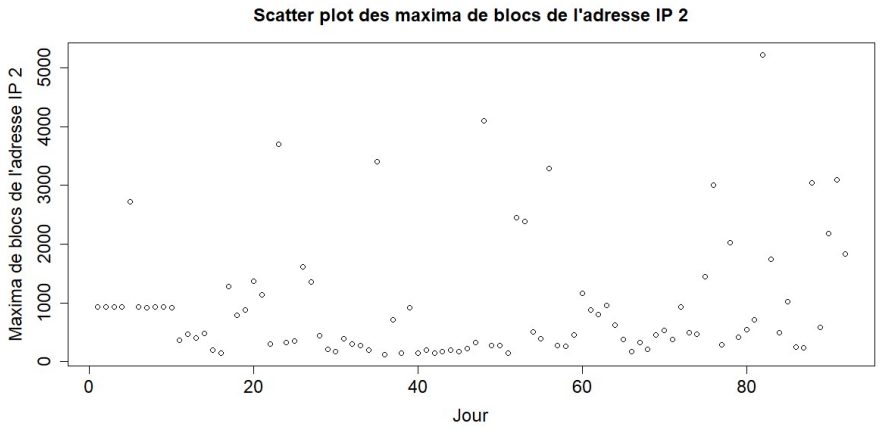


Figure 14: Représentation des maxima de blocs de la variable flux pour l'adresse IP 2 en fonction du jour. Les points représentent les valeurs maximales retenues en tant que maxima de blocs pour chaque jour de la période considérée.

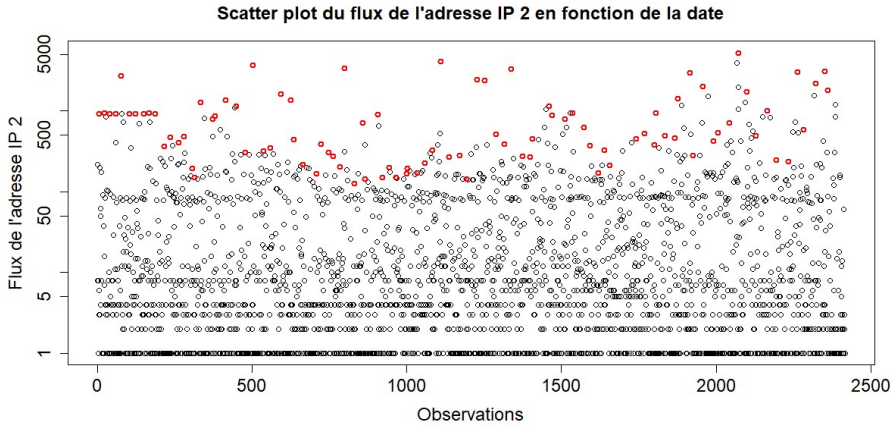


Figure 15: Représentation de la variable flux pour l'adresse IP 2 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique et avec les maxima de blocs indiqués en rouge

Cependant, malgré la mise en évidence des maxima de blocs, il reste difficile de visualiser le caractère maximal de certaines de ces valeurs ainsi que leur appartenance à un bloc particulier. Pour ce faire, les différents blocs peuvent être représentés. La **Figure 16** montre cette représentation. Sur celle-ci, les lignes bleues délimitent les blocs qui ont une longueur égale à un jour. D'ailleurs, nous pouvons remarquer sur ce graphe que les blocs sont de longueurs différentes en termes de nombre total de connexions (flux) en une journée. Cela met en lumière ce que nous expliquions précédemment concernant la taille des blocs.

Toutefois, si une entreprise dispose d'un nombre plus élevé de données, c'est-à-dire un nombre journalier de connexions plus grand, cette dernière peut également décider de prendre des blocs avec une longueur plus large, 2 jours par exemple. Cette configuration peut tout aussi bien être représentée (**Figure 17**). Cette nouvelle configuration est davantage lisible par rapport à la précédente. En revanche, dans le cas présent, une longueur de blocs égale à 2 signifie un échantillon de 46 données. Mais les lecteurs savent désormais qu'il est possible de couper les données en blocs avec la longueur qu'ils désirent en termes d'unité

de temps. En fonction des dimensions de l'ensemble de données, il pourrait être pertinent de prendre des blocs de données par mois ou même par année.

Après avoir défini la taille des blocs et calculé les maxima de blocs, la prochaine étape de la méthode consiste à calculer les estimateurs de maximum de vraisemblance des trois paramètres des maxima de blocs. Pour rappel, ce sont les paramètres de localisation, d'échelle et de forme. Les résultats obtenus sont les suivants. Le paramètre de localisation μ vaut 373,262 avec un intervalle de confiance de $[298{,}120\,;448{,}404]$ et un écart-type de 38,338, le paramètre d'échelle σ vaut 296,699 avec un intervalle de confiance de $[208{,}978\,;384{,}420]$ et un écart-type de 44,756 et le paramètre de forme ξ vaut 0,852 avec un intervalle de confiance de $[0{,}539\,;1{,}165]$ et un écart-type de 0,160.

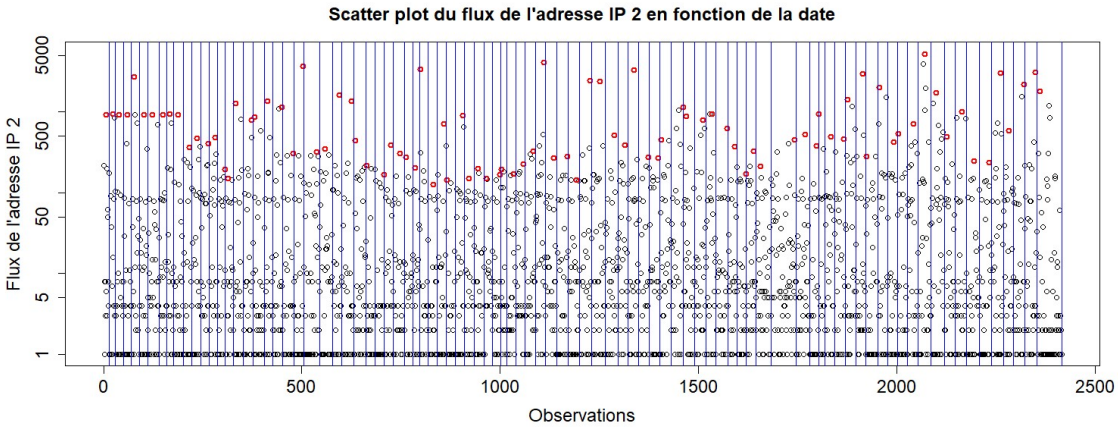


Figure 16: Représentation de la variable flux pour l'adresse IP 2 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique, les maxima de blocs indiqués en rouge et les différents blocs avec une longueur égale à 1 jour délimités par les lignes bleues

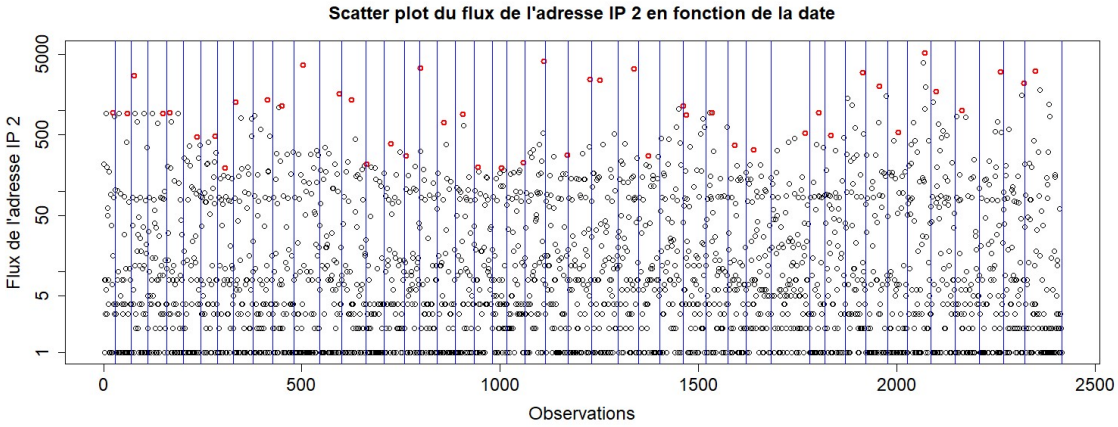


Figure 17: Représentation de la variable flux pour l'adresse IP 2 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique, les maxima de blocs indiqués en rouge et les différents blocs avec une longueur égale à 2 jours délimités par les lignes bleues

Avant de mesurer le niveau grâce à ces estimateurs, un diagnostic du modèle est crucial. Cela va permettre de mesurer et se rendre compte de la qualité de l'ajustement du modèle estimé. La **Figure 18** nous donne les quatre diagrammes pour le diagnostic du modèle,

comme discuté dans la section 4.b.ii.. De manière générale, nous pouvons en conclure que le modèle s'ajuste plutôt bien à nos données. En effet, pour la majorité des données, nous pouvons constater que l'adaptation aux données est de bonne qualité. Cependant, en ce qui concerne les données les plus extrêmes, l'ajustement du modèle est pauvre. Étant donné qu'il y a peu de données avec de grandes valeurs, il est plus difficile de modéliser correctement un modèle qui s'ajustera au mieux à ces données. L'insécurité est donc plus grande à ce niveau-là.

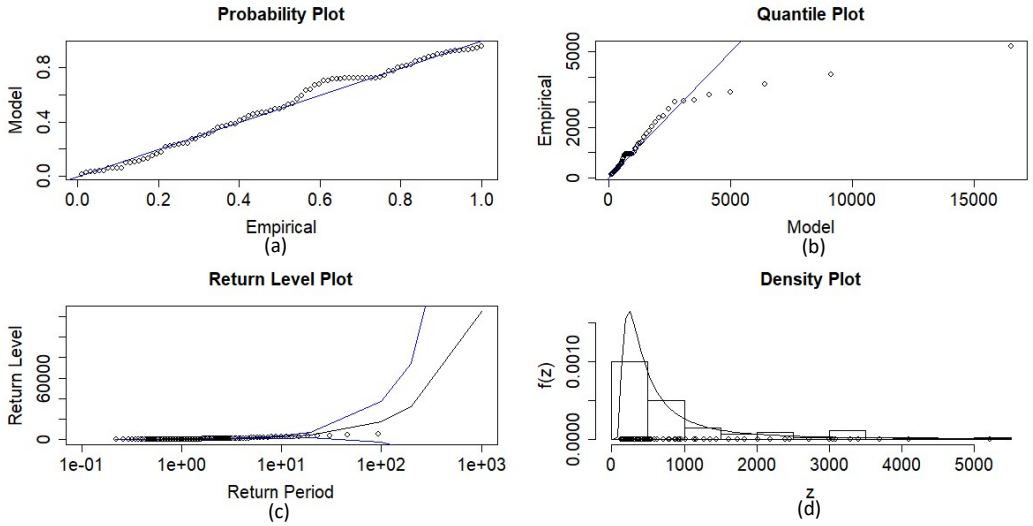


Figure 18: Diagnostic de la qualité de l'ajustement du modèle estimé pour l'adresse IP 2 pour la méthode des maxima de blocs avec les diagrammes de (a) probabilité, (b) de quantile, (c) des niveaux et (d) de densité

Par conséquent, la bonne qualité de l'ajustement du modèle estimé signifie que nous pouvons désormais nous baser sur ce modèle estimé ainsi que sur les paramètres estimés pour prévoir le futur et effectuer des suppositions quant aux flux potentiels futurs d'un site Internet ou d'un système.

Après avoir estimé le modèle et prouvé sa bonne qualité de l'ajustement, il va maintenant être possible de répondre à notre question de recherche. Pour rappel, deux scénarios sont possibles. 'Étant donné un flux, quelle est la probabilité de le dépasser?' est le premier scénario alors que le second est 'Étant donné un niveau de risque, quel est le flux potentiel auquel une entreprise peut s'attendre à atteindre ou à dépasser?'.

1er cas

Comme vu précédemment dans la théorie (section 4.b.ii.), le niveau z_p est tel que $G(z_p) = 1 - p$, c'est-à-dire la probabilité qu'un maximum de bloc le dépasse est égale à p, $P[M > z_p] = p$. Nous allons donc nous baser sur cette formule pour répondre à la question 'Étant donné un certain flux, quelle est la probabilité de le dépasser?'. Plusieurs niveaux de flux vont être pris en considération pour mieux comprendre comment interpréter les résultats obtenus. Les estimateurs de maximum de vraisemblance vont nous être utiles pour calculer les risques liés aux niveaux de flux considérés.

Étant donné que le but de la théorie des valeurs extrêmes est d'estimer quelque chose, dans notre cas un flux, de plus extrême que ce que nous avons déjà observé et obtenu, nous allons dès lors sélectionner deux flux nous permettant d'aller dans ce sens, à savoir un flux de 7 500 et un flux de 75 000. D'ailleurs, un flux trop petit peut être estimé par les probabilités empiriques, ce qui n'est pas ce que nous désirons dans le cas de figure actuel. Nous cherchons en effet à estimer des flux plus extrêmes que ce qui a déjà été observé auparavant et, par conséquent, impossible à mesurer avec les probabilités empiriques.

- $f = 7\,500$

Si une entreprise fait face à un flux de 7 500, la probabilité de dépasser ce flux au moins une fois sur une journée donnée est de 0,02696, soit une probabilité de 2,696%. Pour un flux de cette envergure, une entreprise est consciente qu'elle s'expose à un risque de 2,696%. Son site Internet ou son système a une probabilité quotidienne de 2,696% de ne plus fonctionner si 7 500 est sa capacité maximale.

- $f = 75\,000$

Si une entreprise engendre un flux égal à 75 000, la probabilité d'excéder un flux de ce niveau au moins une fois sur une journée donnée est de 0,00182, à savoir une probabilité de 0,182%. Dans ce cas, le risque encouru par l'entreprise est de moins de 1%. Le risque d'un arrêt quotidien du site ou du système est donc de 0,182% si 75 000 est sa capacité maximale.

Si nous comparons ces deux situations, nous pouvons remarquer que, pour un flux dix fois plus élevé, le risque est 14,813 fois plus faible.

Toutefois, en vue d'être performantes, les entreprises doivent effectuer ce calcul pour un flux f correspondant à leur capacité maximale. De cette manière, elles pourront déterminer le risque encouru d'un dysfonctionnement de leur site ou de leur système avec une telle capacité et adapter leurs ressources si nécessaire. Par exemple, si une entreprise désire prendre le moins de risque possible, cette dernière devra allouer davantage de budget à la gestion de son site Internet ou de son système. En effet, au plus le budget concédé à la gestion du site Internet ou du système est large, au plus la capacité de ce dernier sera élevée et au plus le risque de panne sera faible. À nouveau, certaines entreprises seront plus enclines à encourir des risques plus importants que d'autres en raison de leur secteur d'activité. Cependant, certaines entreprises n'auront d'autres choix que de s'exposer à un grand risque en raison d'un manque de moyens.

En résumé, cette technique permettra aux entreprises de connaître le risque potentiel pour un flux donné. De manière plus précise, les entreprises seront capables, grâce à cette méthode, d'estimer le risque pour des flux extrêmes qu'elles ne sont pas capables de gérer. Cette technique se voit donc être un énorme avantage pour les entreprises qui pourront dès lors s'organiser afin de gérer au mieux leur site Internet et/ou leurs systèmes.

2ème cas

Pour le deuxième scénario, le niveau z_p doit être calculé. Ce niveau pour les maxima de blocs est donné par l'équation 2 dans la section 4.b.ii.. Les trois paramètres estimés par le maximum de vraisemblance vont nous être utiles pour mesurer ce niveau. De plus, pour calculer ce niveau, nous devons fixer p, c'est-à-dire fixer le risque qu'une entreprise est prête à prendre quant à la possibilité d'un arrêt de son site ou de son système. De manière générale, les entreprises préfèreront un niveau de risque faible. En vue d'illustrer cette pratique, nous allons opter pour divers niveaux de risque, à savoir 1%, 0,2739% et 0,06845%.

- $p = 1\% = 0{,}01$

Si $p = 0{,}01$, le niveau $z_{0{,}01}$ vaut 17 553,260. Cela signifie que 17 553,260 est le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,01. En moyenne, cela devrait arriver tous les 100 jours $\left(= {}^{1}/_{0{,}01}\right)$. Le niveau $z_{0{,}01}$ a un intervalle de confiance de $[-2\,168{,}743\,;\,37\,275{,}270]$ avec un écart-type de 10 062,250.

- $p = 0{,}2739\% = 0{,}002739$

Si $p = 0{,}002739$, le niveau $z_{0{,}002739}$ vaut 53 010,410. Cela signifie que 53 010,410 est le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,002739. En moyenne, cela devrait arriver tous les 365 jours, soit tous les ans $\left(= {}^{1}/_{0{,}002739}\right)$. Le niveau $z_{0{,}002739}$ a un intervalle de confiance de $[-27\,415{,}880\,;\,133\,436{,}700]$ avec un écart-type de 41 033,830.

- $p = 0{,}06845\% = 0{,}0006845$

Si $p = 0{,}0006845$, le niveau $z_{0{,}0006845}$ vaut 172 813,400. Cela signifie que 172 813,400 est le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,0006845. En moyenne, cela devrait arriver tous les 1 461 jours, soit tous les quatre ans $\left(= {}^{1}/_{0{,}0006845}\right)$. Le niveau $z_{0{,}0006845}$ a un intervalle de confiance de $[-163\,199{,}800\,;\,508\,826{,}700]$ avec un écart-type de 171 435,300.

De manière générale, nous pouvons constater que les intervalles de confiance pour les niveaux sont très larges. Cela est tout à fait normal vu la taille de notre échantillon. Une meilleure précision pourrait être obtenue si nous avions un ensemble de données plus grand. Cependant, les entreprises ont généralement des données pour une période bien plus large que trois mois.

En outre, le niveau de risque sélectionné dépendra bien évidemment de la nature de l'entreprise en question. Une entreprise dans le secteur boursier voudra s'assurer d'un fonctionnement durable de son site ou de ses systèmes en raison de la nature extrêmement

sensible de leur activité. En effet, chaque seconde, les données changent. Une panne des systèmes serait catastrophique. La conséquence d'un arrêt des systèmes pourrait leur faire perdre l'opportunité de gagner des millions d'euros voire bien plus. Elles choisiront dès lors un risque très faible. En revanche, une entreprise offrant des vêtements en ligne peut se permettre un risque plus élevé puisque les conséquences sont moins catastrophiques. Le risque fixé sera dès lors plus élevé.

Tout dépend non seulement de la nature mais aussi de la taille de l'entreprise. Un faible risque sera aussi de manière assez globale choisi par les grandes entreprises. En effet, ces dernières auront besoin d'une capacité plus grande étant donné le nombre plus large de visiteurs sur leur site. En ce qui concerne les petites entreprises avec moins de clientèle, elles choisiront un risque plus élevé. En définitive, le risque encouru dépend entièrement de l'importance de la conséquence pour l'entreprise dans le cas d'un arrêt du site ou système.

De plus, le niveau (return level) est le plus souvent ce que les entreprises recherchent à savoir. En effet, grâce à cette information, les entreprises pourront prendre des décisions quant à la gestion de leur site Internet et de leur système. Par exemple, une entreprise se rend compte grâce à cette méthode qu'un potentiel arrêt de son site ou système se produira tous les ans. Mais, en réalité, son objectif est d'avoir un potentiel arrêt tous les cinq ans. En vue de régler ce souci et d'atteindre cette fin, celle-ci devra allouer davantage de ressources. Il se peut également qu'une entreprise réalise qu'elle alloue assez bien ses ressources pour le bon fonctionnement du système. Dans ce cas, elle décidera de continuer de cette manière et allouera les ressources supplémentaires prévues pour le maintien du système à un autre domaine de son activité. En conséquence, la gestion globale de l'entreprise s'en verra améliorer, maximiser.

b. Adresse IP 3

Nous allons maintenant nous intéresser à l'adresse IP 3 au regard de la méthode des maxima de blocs. L'approche reste identique. Premièrement, la taille des blocs est choisie. À nouveau, la taille des blocs est fixée à une longueur égale à un jour. Ensuite, les maxima de blocs sont générés et représentés. La **Figure 19** représente ces maxima de blocs pour l'adresse IP 3. D'ailleurs, une petite précision est nécessaire. Pour l'adresse IP 3, le nombre de maxima de blocs s'élève à 89 au lieu de 92. Cela s'explique par le fait qu'aucun flux ne s'est produit pour trois jours particuliers. Ces derniers sont les trois derniers jours, soit le 28-29-30 Septembre. Ensuite, tout comme pour l'adresse IP 2, nous pouvons représenter les maxima en fonction du flux global de l'adresse IP 3 (**Figure 20**) mais aussi présenter un graphe reprenant les divers blocs avec leur maximum (**Figure 21**).

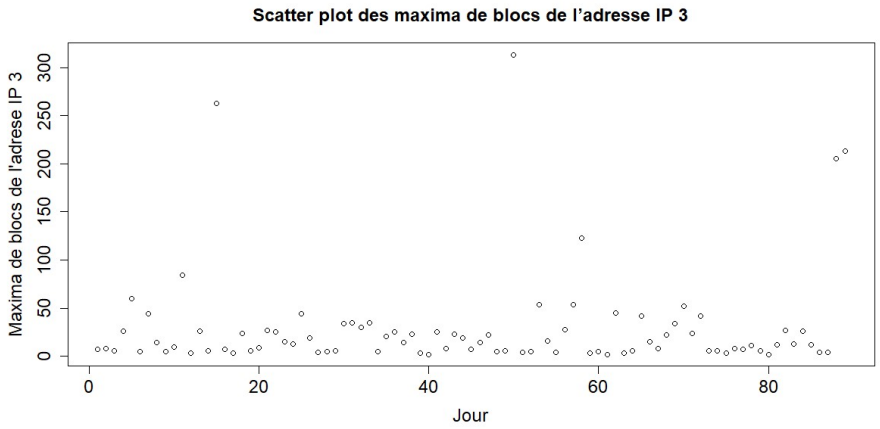


Figure 19: Représentation des maxima de blocs de la variable flux pour l'adresse IP 3 en fonction du jour. Les points représentent les valeurs maximales retenues en tant que maxima de blocs pour chaque jour de la période considérée.

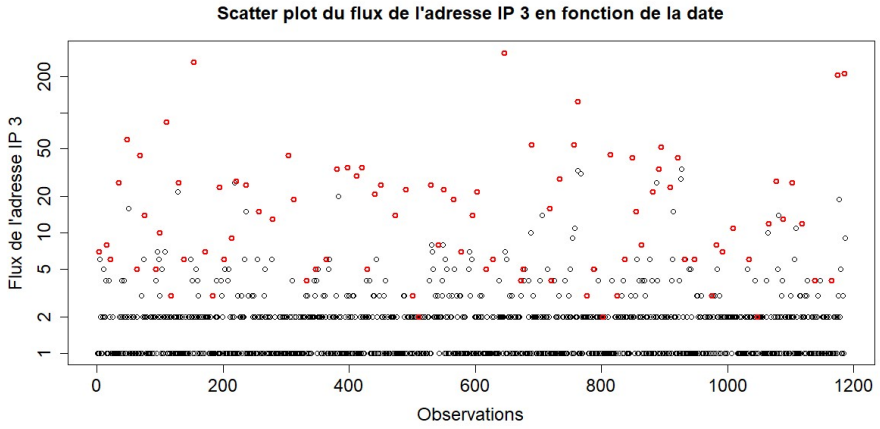


Figure 20: Représentation de la variable flux pour l'adresse IP 3 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique et avec les maxima de blocs indiqués en rouge

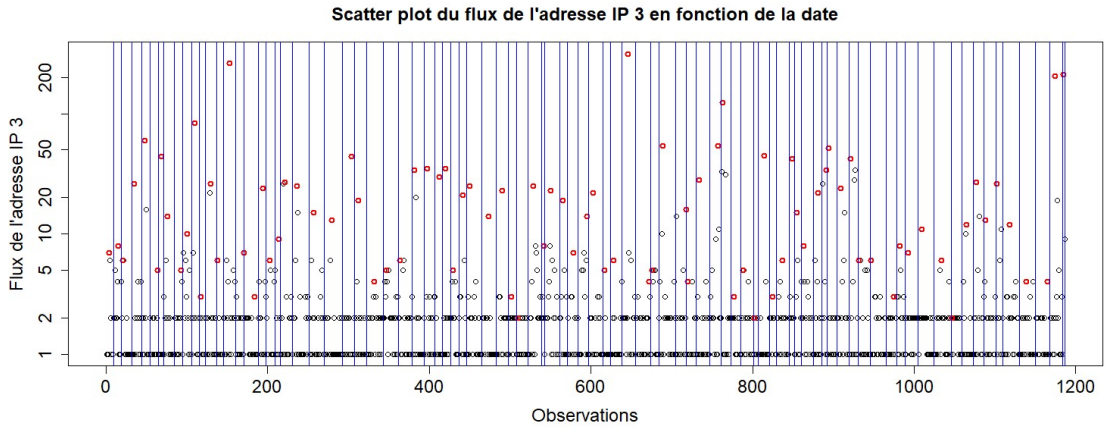


Figure 21: Représentation de la variable flux pour l'adresse IP 3 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique, les maxima de blocs indiqués en rouge et les différents blocs avec une longueur égale à 1 jour délimités par les lignes bleues

Ensuite vient le moment d'estimer les paramètres grâce à la technique du maximum de vraisemblance. Pour l'adresse IP 3, le paramètre de localisation μ a une valeur de 7,964 avec un intervalle de confiance de $[6{,}112\,;9{,}816]$ et un écart-type de 0,945. Le paramètre d'échelle σ vaut 7,578 avec un intervalle de confiance de $[5{,}205\,;9{,}951]$ et un écart-type de 1,211. Le paramètre de forme ξ a une valeur de 0,963 avec un intervalle de confiance de $[0{,}670\,;1{,}256]$ et un écart-type de 0,149. D'ailleurs, le paramètre de forme ξ de l'adresse IP 3 est un peu plus élevé que celui de l'adresse IP2, ce qui signifie des queues de distribution un peu plus longues pour l'adresse IP 3.

Enfin, l'évaluation de la qualité de l'ajustement du modèle estimé doit être réalisée. La **Figure 22** nous fait remarquer que le modèle s'ajuste assez bien aux données. Le diagramme des quantiles montre une seule observation (la plus grande) qui dévie plus fortement de la droite.

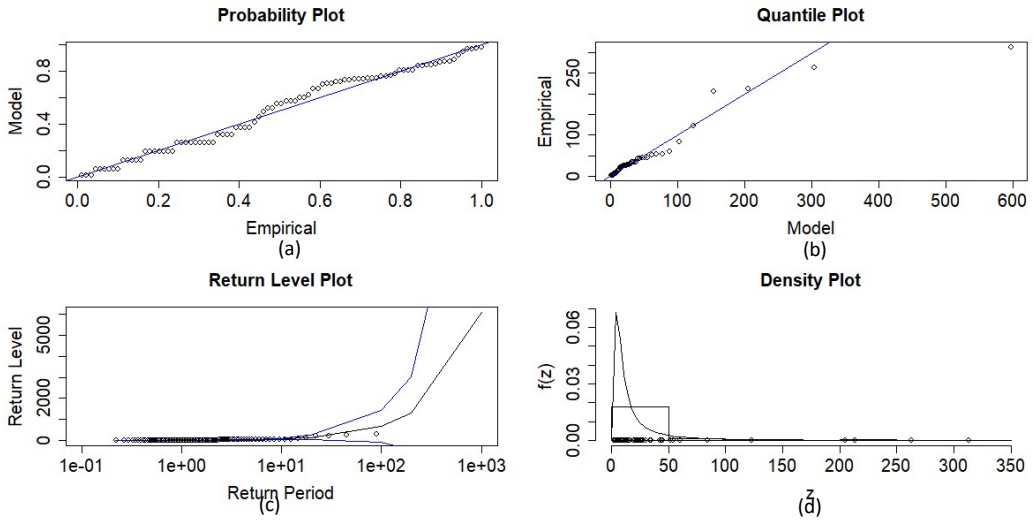


Figure 22: Diagnostic de la qualité de l'ajustement du modèle estimé pour l'adresse IP 3 pour la méthode des maxima de blocs avec les diagrammes de (a) probabilité, (b) de quantile, (c) des niveaux et (d) de densité

À nouveau, une fois toutes ces étapes terminées, nous pouvons passer à la question de recherche. Nous allons encore une fois répondre à cette problématique pour les deux scénarios envisagés.

1er cas

De la même manière que pour l'adresse IP 2, nous allons appliquer la formule $P[M > z_p] = p$ afin de déterminer la probabilité qu'un maximum de blocs dépasse un flux donné. Plusieurs niveaux de flux vont être sélectionnés dans le but de comprendre le fonctionnement de cette pratique et d'interpréter les résultats générés. Les estimateurs de maximum de vraisemblance vont une fois encore être utilisés afin de mesurer le risque correspondant. Deux flux avec des envergures différentes, 600 et 6 000, seront choisis afin d'illustrer le principe d'estimation de flux plus extrêmes que ce que nous avons observé dans le passé, ce qui constitue l'objectif de la théorie des valeurs extrêmes.

- $f = 600$

La probabilité de dépasser un flux de 600 au moins une fois sur une journée donnée est de 0,01104, à savoir une probabilité de 1,104%. Le risque encouru par l'entreprise est dès lors d'environ 1%, ce qui représente le pourcentage associé à la panne potentielle quotidienne du site ou du système si 600 est sa capacité maximale.

- $f = 6\,000$

La probabilité de surpasser un flux de 6 000 au moins une fois sur une journée donnée est de 0,00102, soit une probabilité de 0,102%. Si une entreprise fait face à un flux identique, elle est consciente qu'elle s'expose à un risque d'environ 0,1%. Son site ou système a un risque de 0,102% de s'arrêter de manière quotidienne si 6 000 est sa capacité maximale.

Si nous comparons ces deux contextes, nous pouvons remarquer que, pour un flux dix fois plus élevé, le risque est 10,823 fois plus faible.

2ème cas

En ce qui concerne le deuxième scénario, les estimateurs du maximum de vraisemblance vont à nouveau nous aider à calculer le niveau. Comme mentionné précédemment, p doit être déterminé avant tout autre chose. En vue d'être cohérent, nous allons adopter les mêmes niveaux de risque que ceux utilisés pour l'adresse IP 2. De cette manière, une comparaison entre les résultats des deux adresses IP pourra être effectuée.

- $p = 1\% = 0{,}01$

Si $p = 0{,}01$, le niveau $z_{0{,}01}$ vaut 660,234, à savoir le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les 100 jours. Le niveau $z_{0{,}01}$ a un écart-type de 386,116 et donc un intervalle de confiance de $[-96{,}553\ ;\ 1\,417{,}021]$.

- $p = 0{,}2739\% = 0{,}002739$

Si $p = 0{,}002739$, le niveau $z_{0{,}002739}$ vaut 2 305,221, à savoir le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les ans. Le niveau $z_{0{,}002739}$ a un écart-type de 1 781,445 et donc un intervalle de confiance de $[-1\,186{,}412\ ;\ 5\,796{,}853]$.

- $p = 0{,}06845\% = 0{,}0006845$

Si $p = 0{,}0006845$, le niveau $z_{0{,}0006845}$ vaut 8 769,912, à savoir le flux que nous pensons atteindre ou dépasser au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les quatre ans. Le niveau $z_{0{,}0006845}$ a un intervalle de confiance de $[-8\,014{,}371\ ;\ 25\,554{,}190]$ avec un écart-type de 8 563,409.

Puis, en ce qui concerne les niveaux, l'explication présentée précédemment pour l'adresse IP 2 tient également pour l'adresse IP 3, c'est-à-dire que la nature et la taille d'une entreprise déterminent le niveau de risque qu'elle est prête à prendre.

Pour conclure cette section, après évaluation des deux adresses IP d'intérêt, nous pouvons constater que cette méthode des maxima de blocs fonctionne pour ces deux adresses IP. Par conséquent, cette technique s'avère finalement être valide, peu importe la quantité de flux prise en considération.

2. Dépassements de seuil

Nous allons maintenant passer à la seconde méthode de la théorie des valeurs extrêmes, à savoir la méthode des dépassements de seuil. Nous allons à nouveau commencer par l'adresse IP 2 et, ensuite, nous nous concentrerons sur l'adresse IP 3.

a. Adresse IP 2

Comme découvert dans la partie théorique (section 4.b.iii.), la première étape consiste à déterminer le seuil au-dessus duquel les valeurs seront considérées comme extrêmes. Cette détermination du seuil s'effectue sur base d'un graphique. Pour rappel, il faut choisir le seuil pour lequel les paramètres d'échelle et de forme sont assez stables. La variabilité de ces paramètres estimés est déterminée par les barres verticales illustrées sur le graphique. Au plus la barre est grande, au plus la variabilité est grande. La **Figure 23** va nous aider à estimer le seuil pour l'adresse IP 2. Nous pouvons remarquer que 400 semble être un seuil raisonnable. Le paramètre d'échelle est relativement constant et la variation du paramètre de forme est acceptable. Cependant, une vraie stabilité n'existe pas pour ces données. En effet, un saut peut être perçu aux alentours de la valeur 900, que nous ne choisissons pas parce que nous aurions très peu de données en notre possession. De plus, ce graphe peut nous donner l'impression qu'il semblerait y avoir plusieurs distributions adaptées à ce jeu de données, à savoir une pour les valeurs moyennes, une pour les valeurs modérément extrêmes et une pour les valeurs très extrêmes.

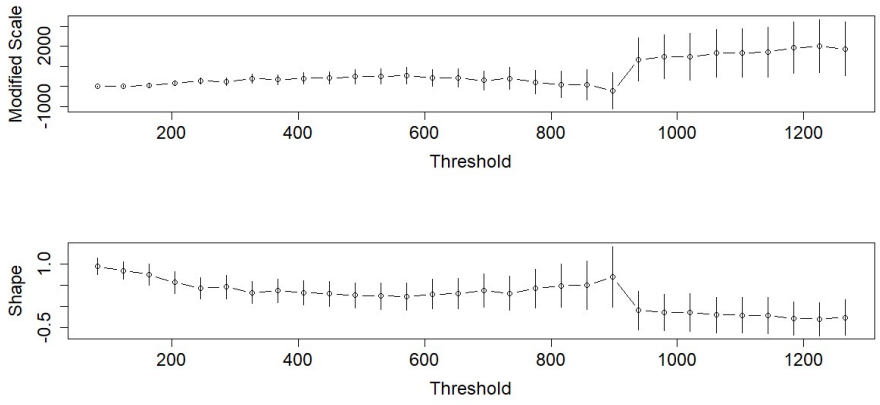


Figure 23: Diagramme de détermination du seuil pour l'adresse IP 2

Après avoir fixé le seuil, le nombre de données étant considérées comme extrêmes peut être découvert. Pour un seuil fixé à 400, le nombre s'élève à 99. Ce dernier est presque identique au nombre de données prises en considération pour la méthode des maxima de blocs, à savoir 92. Par conséquent, il sera possible de comparer les méthodes en termes de variabilité. Ces valeurs jugées comme extrêmes peuvent être représentées en fonction du flux global de l'adresse IP 2 (**Figure 24**).

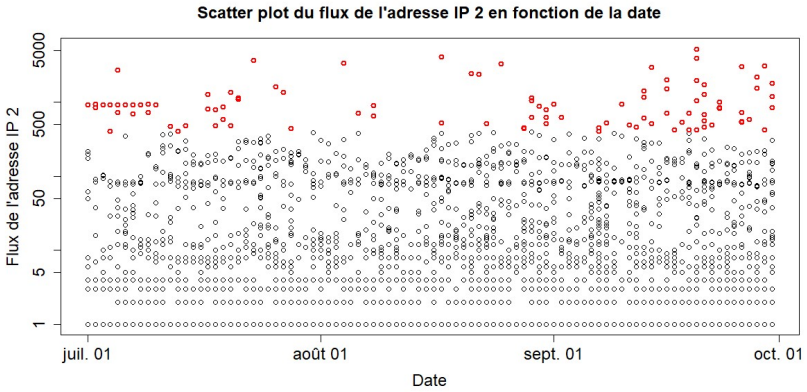


Figure 24: Représentation de la variable flux pour l'adresse IP 2 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique et avec les valeurs dépassant le seuil de 400 indiquées en rouge (valeurs extrêmes)

Ensuite, nous pouvons établir avec quelle probabilité le seuil égal à 400 est dépassé. Cette probabilité est de 4%, ce qui signifie que 96% des données possèdent une valeur inférieure ou égale à 400. Puis, l'étape suivante est d'estimer les divers paramètres, à savoir les paramètres d'échelle σ et de forme ξ, grâce au maximum de vraisemblance. Pour la méthode des dépassements de seuil, ces paramètres vont être estimés en fonction du seuil choisi. Les estimations obtenues sont les suivantes. Le paramètre d'échelle σ vaut 539,790 avec un intervalle de confiance de $[358{,}437\,;721{,}143]$ et un écart-type de 92,527. Le paramètre de forme ξ, quant à lui, a une valeur de 0,302 avec un intervalle de confiance de $[0{,}024\,;0{,}581]$ et un écart-type de 0,142.

Avant de passer à la dernière étape, une comparaison entre les paramètres de forme ξ de la méthode des maxima de blocs et celle des dépassements de seuil peut nous être utile. Pour rappel, le paramètre de forme des maxima de blocs était de 0,852 avec un intervalle de confiance de $[0{,}539\,;1{,}165]$ alors que celui des dépassements de seuil vaut 0,302 avec un intervalle de confiance de $[0{,}024\,;0{,}581]$. Sur base de ces informations, nous pouvons voir que les intervalles de confiance ne se chevauchent que pour très peu de données, entre 0,539 et 0,581. Les dépassements de seuil n'ont donc pas le même comportement que les maxima de blocs. De plus, les deux ξ sont positifs. Cela signifie que nous nous situons dans un domaine de Fréchet. Autrement dit, les extrémités seront très longues. Plus ξ est grand, plus l'extrémité droite de la distribution est longue et plus il y a de très grands extrêmes.

Enfin, la dernière étape consiste à évaluer la qualité de l'ajustement du modèle estimé. Sur base de la **Figure 25**, il est possible de conclure que le modèle est bien ajusté à nos données.

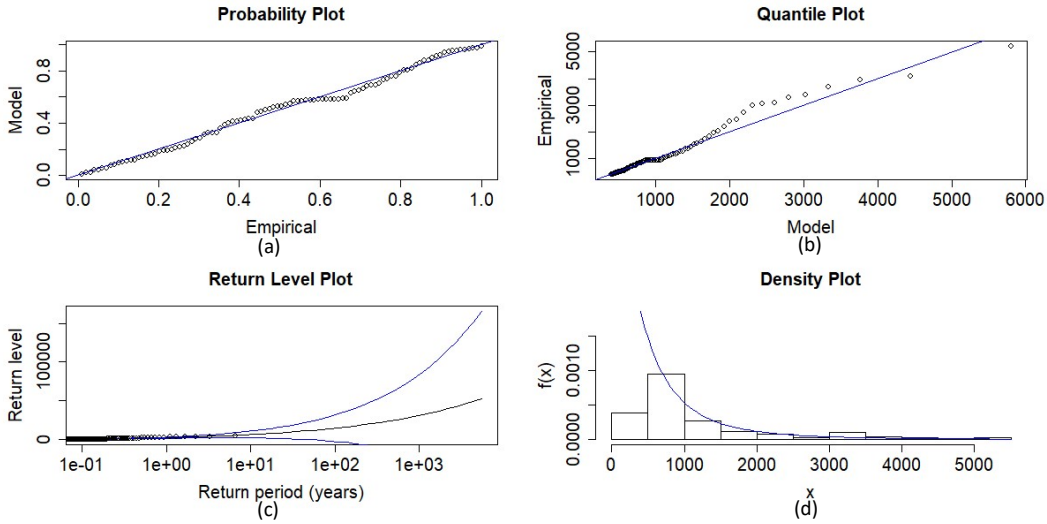


Figure 25: Diagnostic de la qualité de l'ajustement du modèle estimé pour l'adresse IP 2 pour la méthode des dépassements de seuil avec les diagrammes de (a) probabilité, (b) de quantile, (c) des niveaux et (d) de densité

Une fois le modèle estimé et la bonne qualité de l'ajustement démontrée, il est désormais temps de répondre à notre question de recherche. Les deux scénarios vont à nouveau être pris en compte. Le premier consiste à déterminer le risque encouru par les entreprises si elles font face à un certain niveau de flux. Le deuxième scénario, quant à lui, rend possible l'estimation du niveau de flux attendu sur base d'un niveau de risque donné.

1er cas

Pour la technique des dépassements de seuil, l'estimation de la probabilité d'un certain flux est tout à fait réalisable. La détermination de cette probabilité se base sur l'équation 5. Les estimateurs de maximum de vraisemblance des deux paramètres, les paramètres d'échelle et de forme, vont être utilisés afin d'évaluer ces probabilités. Pour l'adresse IP 2, le seuil est fixé à 400. Étant donné que ce seuil est choisi par l'entreprise, celle-ci s'intéressera donc à des flux supérieurs à ce dernier, mais surtout plus extrêmes que ce qu'elle a déjà observé. Divers niveaux de flux vont être choisis pour assimiler le fonctionnement et interpréter les résultats recueillis, à savoir un flux de 7 500 et un flux de 75 000, comme pour l'adresse IP 2, afin de rester cohérent dans nos choix.

- $f = 7\,500$

Si une entreprise fait face à un flux équivalent à 7 500, cette dernière aura une probabilité de 0,0002028, soit une probabilité de 0,02028%, de dépasser un flux valant 7 500. En conséquence, une entreprise prendrait un risque de moins de 1% de voir son site Internet ou son système ne plus fonctionner si 7 500 est sa capacité maximale.

- $f = 75\,000$

La probabilité d'excéder un flux de 75 000 pour une entreprise générant un flux de cette envergure est de $1{,}643 * 10^{-7}$, soit une probabilité de 0,00001643%. Une entreprise

encourt un risque très faible de 0,00001643% d'excéder un flux de 75 000 si 75 000 est sa capacité maximale.

Si nous comparons ces deux situations, nous pouvons remarquer que, pour un flux dix fois plus élevé, le risque est 1 234,327 fois plus faible.

À nouveau, cette technique permet aux entreprises d'estimer le risque pour des flux qu'elles estiment pouvant apparaître à l'avenir, mais surtout estimer le risque pour des flux plus extrêmes qu'elles ne sont capables de gérer. Cela représente un avantage significatif pour les entreprises qui pourront dès lors gérer au mieux leur site Internet et/ou leurs systèmes.

Cependant, une constatation est que les niveaux de risque sont relativement petits comparés à ceux de la méthode des maxima de blocs. Pour un flux de 7 500, le risque est de 0,02028% contre un risque de 2,696% et, pour un flux de 75 000, le risque vaut 0,00001643% contre un risque de 0,182%. L'explication réside dans l'élément temporel. En effet, la méthode des maxima de blocs va considérer le risque de dépasser un certain niveau de flux au moins une fois sur une journée donnée alors que la méthode des dépassements de seuil va plutôt considérer la probabilité de surpasser un flux de manière globale, à n'importe quel moment. C'est pourquoi, les probabilités des maxima de blocs sont plus élevées que celles des dépassements de seuil.

2ème cas

En ce qui concerne le deuxième scénario, un niveau peut également être mesuré pour les dépassements de seuil. Dans cette technique, le niveau est déterminé par x_m, m représentant le produit entre l'étendue de la période pour laquelle nous nous intéressons, N, et le nombre d'observations moyen par jour, n_y.

Dans la méthode des maxima de blocs, le flux potentiel avec une probabilité p pouvait se produire tous les $\frac{1}{p}$ jours. Pour les dépassements de seuil, la probabilité est déterminée par $\frac{1}{N}$, c'est-à-dire que $p = \frac{1}{N}$. Cela signifie que, si nous connaissons la probabilité, nous connaitrons également la période pendant laquelle cela se produira. En vue de mettre en lumière cette méthode, nous allons sélectionner divers niveaux de risque, à savoir 1%, 0,2739% et 0,06845%.

- $p = 1\% = 0{,}01$

Si $p = 0{,}01$, le niveau x_m vaut 5 958,825. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,01. En moyenne, cela devrait arriver tous les 100 jours $\left(= {}^{1}/_{0{,}01}\right)$. Le niveau x_m a un intervalle de confiance de $[567{,}382\ ;\ 11\,350{,}270]$ avec un écart-type de 2 750,736.

- $p = 0{,}2739\% = 0{,}002739$

Si $p = 0{,}002739$, le niveau x_m vaut 9 476,471. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,002739. En moyenne, cela devrait arriver tous les 365 jours, soit tous les ans $\left(= {}^{1}/_{0{,}002739}\right)$. Le niveau x_m a un intervalle de confiance de $[-570{,}706\,;\,19\,523{,}650]$ avec un écart-type de 5 126,111.

- $p = 0{,}06845\% = 0{,}0006845$

Si $p = 0{,}0006845$, le niveau x_m vaut 15 133,050. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée avec une probabilité de 0,0006845. En moyenne, cela devrait arriver tous les 1 461 jours, soit tous les quatre ans $\left(= {}^{1}/_{0{,}0006845}\right)$. Le niveau x_m a un écart-type de 10 041,91 et un intervalle de confiance de $[-4\,549{,}099\,;\,34\,815{,}200]$.

À présent, nous allons comparer les flux obtenus pour les deux méthodes d'intérêt, la méthode des maxima de blocs et des dépassements de seuil pour les trois niveaux de risque considérés. Pour un risque de 1%, le flux potentiel est de 5 958,825 contre 17 553,260. Pour un risque de 0,2739%, le flux se situe aux alentours de 9 500 contre 53 000. Enfin, pour un risque de 0,06845%, le flux vaut 15 133,050 contre 172 813,400. À nouveau, cela est dû au fait que les maxima de blocs travaillent sur base du quotidien alors que les dépassements de seuil travaillent sur la période entière. Nous nous attendons donc à avoir des flux plus élevés pour les maxima de blocs que pour les dépassements de seuil et il s'est avéré que cela est effectivement le cas. Pour un même risque, le niveau de flux nécessaire pour dépasser un seuil est plus élevé pour les maxima de blocs que pour les dépassements de seuil.

b. Adresse IP 3

Maintenant, nous allons nous concentrer sur l'adresse IP 3. Les mêmes étapes vont lui être appliquées. À nouveau, la première étape est de sélectionner un seuil adéquat à partir du graphique dédié à cela (**Figure 26**). Dans ce cas-ci, la stabilité des paramètres est très bonne. La valeur 20 semble être un bon compromis. Ce seuil de 20 engendre dès lors un nombre de 43 valeurs extrêmes, c'est-à-dire que 43 valeurs dépassent le seuil de 20. D'ailleurs, la **Figure 27** expose ces valeurs considérées comme extrêmes par rapport au flux global de l'adresse IP 3. De plus, dans le cas de l'adresse IP 3, ce seuil de 20 correspond à la probabilité de se situer en-dessous de 20. Cette probabilité vaut 96%. Nous avons donc approximativement le même pourcentage de données que celui de l'adresse IP 2.

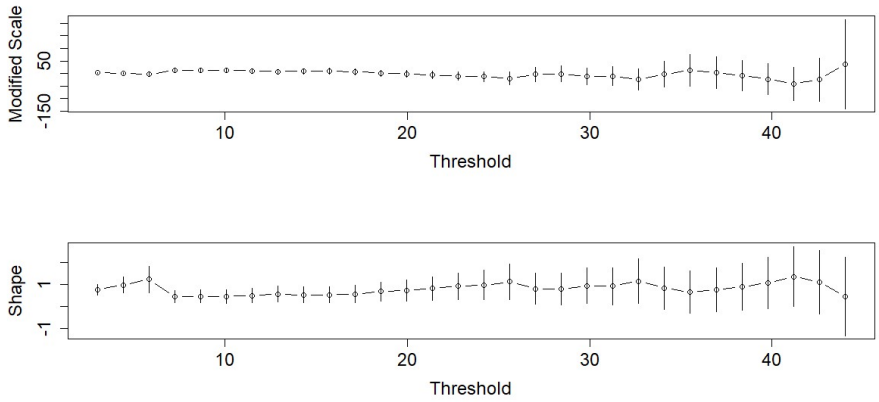


Figure 26: Diagramme de détermination du seuil pour l'adresse IP 3

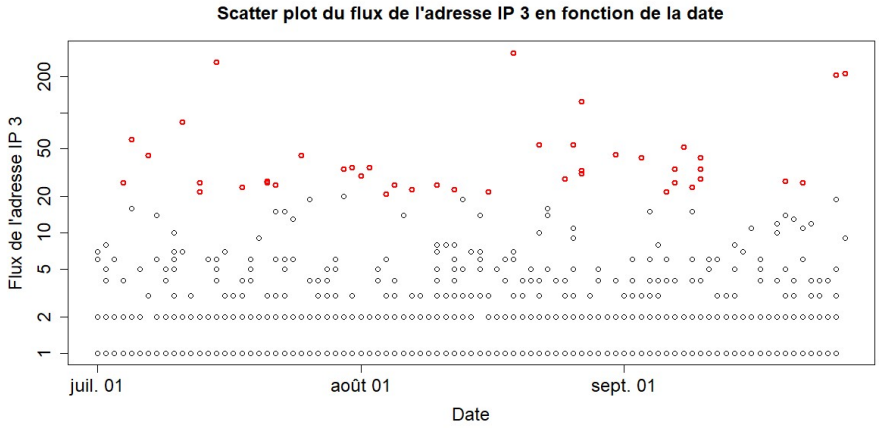


Figure 27: Représentation de la variable flux pour l'adresse IP 3 en fonction de la variable date, avec l'axe des ordonnées en échelle logarithmique et avec les valeurs dépassant le seuil de 14 indiquées en rouge (valeurs extrêmes)

Puis, l'étape suivante est d'estimer les paramètres à l'aide du maximum de vraisemblance. La valeur du paramètre d'échelle σ est de 13,526 avec un intervalle de confiance de $[6{,}628\,;\,20{,}425]$ et un écart-type de 3,520. La valeur du paramètre de forme ξ de 0,699 avec un intervalle de confiance de $[0{,}236\,;\,1{,}161]$ et un écart-type de 0,236. D'ailleurs, le paramètre de forme ξ de l'adresse IP 3 est relativement plus élevé que celui de l'adresse IP2, ce qui signifie des queues de distribution plus longues pour l'adresse IP 3.

Avant de passer à la dernière étape, une comparaison entre les paramètres de forme ξ de la méthode des maxima de blocs et celle des dépassements de seuil peut nous être utile. Pour rappel, le paramètre de forme des maxima de blocs était de 0,963 avec un intervalle de confiance de $[0{,}670\,;\,1{,}256]$ alors que celui des dépassements de seuil vaut 0,699 avec un intervalle de confiance de $[0{,}236\,;\,1{,}161]$. Sur base de ces informations, nous pouvons voir que les intervalles de confiance se chevauchent pour une grande majorité de données, entre 0,670 et 1,161. Les maxima de blocs et les dépassements de seuil ont donc un comportement similaire. Les deux ξ sont aussi positifs. Cela signifie que nous nous situons à nouveau dans un domaine de Fréchet. Autrement dit, les extrémités seront très longues.

Enfin, la qualité de l'ajustement du modèle estimé peut être évaluée. À partir de la **Figure 28**, nous pouvons déduire que ce modèle est très adapté à nos données. Certaines valeurs les plus extrêmes dévient du modèle. Cependant, nous pouvons constater que l'ajustement de l'adresse IP 3 est meilleur que celui de l'adresse IP2 pour une grande partie des données.

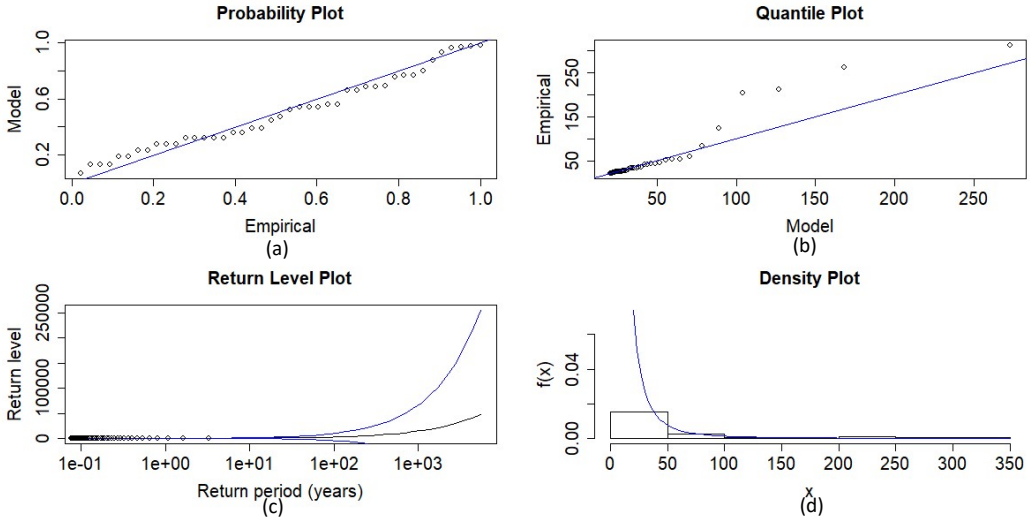


Figure 28: Diagnostic de la qualité de l'ajustement du modèle estimé pour l'adresse IP 3 pour la méthode des dépassements de seuil avec les diagrammes de (a) probabilité, (b) de quantile, (c) des niveaux et (d) de densité

Après avoir estimé le modèle et prouvé la bonne qualité de l'ajustement, nous pouvons répondre à la question de recherche. À nouveau, nous allons considérer les deux scénarios.

1er cas

Pour déterminer la probabilité de dépasser un certain flux, les estimateurs de maximum de vraisemblance sont pris en compte. Pour l'adresse IP 3, le seuil fixé auparavant est de 20. L'intérêt des entreprises est d'évaluer ce risque pour des flux plus extrêmes que ce seuil et que ce qu'elle a déjà observé. Nous allons reprendre les flux choisis dans la méthode des maxima de blocs pour être en accord, à savoir des seuils de 600 et de 6 000.

- $f = 600$

Une entreprise aura une probabilité de 0,000266, soit une probabilité de 0,027%, de dépasser un flux égal à 600. Une entreprise prendrait un risque de 0,027% de voir son site Internet ou son système ne plus fonctionner si 600 est capacité maximale.

- $f = 6\,000$

Une entreprise faisant face à un flux de 6 000 a une probabilité de $9{,}846 * 10^{-6}$, soit une probabilité de 0,000985%, de surpasser un flux de cette envergure. Cette dernière sera donc encline à risquer une panne de son site Internet ou de ses systèmes avec un risque très faible de 0,000985% si 6 000 est sa capacité maximale.

Si nous comparons ces deux contextes, nous pouvons remarquer que, pour un flux dix fois plus élevé, le risque est 27,411 fois plus faible.

Cependant, par rapport à la méthode des maxima de blocs, les niveaux de risque sont relativement petits. Pour un flux de 600, le risque est de 0,027% contre un risque de 1,104% et, pour un flux de 6 000, le risque vaut 0,000985% contre un risque de 0,102%. Encore une fois, l'explication réside dans l'élément temporel. En effet, la méthode des maxima de blocs va considérer le risque de dépasser un certain niveau de flux de manière quotidienne alors que la méthode des dépassements de seuil va plutôt considérer la probabilité de surpasser un flux de manière globale, à n'importe quel moment.

2ème cas

Pour le deuxième scénario, nous allons choisir plusieurs probabilités dans le but d'appliquer cette pratique et d'analyser les résultats générés. Pour rappel, la connaissance de la probabilité permettra de connaître le moment où ce type de flux apparaîtra.

- $p = 1\% = 0{,}01$

Si $p = 0{,}01$, le niveau x_m vaut 291,306. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les 100 jours. Le niveau x_m a un écart-type de 326,122 et un intervalle de confiance de $[-347{,}892\ ;\ 930{,}504]$.

- $p = 0{,}2739\% = 0{,}002739$

Si $p = 0{,}002739$, le niveau x_m vaut 718,746. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les ans. Le niveau x_m a un écart-type de 915,892 et donc un intervalle de confiance de $[-1\,076{,}402\ ;\ 2\,513{,}894]$.

- $p = 0{,}06845\% = 0{,}0006845$

Si $p = 0{,}0006845$, le niveau x_m vaut 1 892,846. Ce nombre représente le flux pouvant être atteint ou dépassé au moins une fois lors d'une certaine journée donnée. En moyenne, cela devrait arriver tous les quatre ans. Le niveau x_m a un écart-type de 2 824,053 et un intervalle de confiance de $[-3\,642{,}297\ ;\ 7\,427{,}990]$.

En outre, l'explication donnée pour l'adresse IP 2 pour le deuxième scénario est aussi valable pour l'adresse IP 3. Si nous comparons les adresses IP 2 et 3 pour la technique des dépassements de seuil, un élément essentiel peut être relevé en ce qui concerne le deuxième scénario. Pour des risques identiques, les flux potentiels générés diffèrent largement. Cet aspect a également été souligné pour la méthode des maxima de blocs.

À présent, les deux méthodes d'intérêt peuvent être comparées en termes de flux obtenus pour les trois niveaux de risque considérés. Pour un risque de 1%, le flux potentiel est de 291,306 contre 660,234. Pour un risque de 0,2739%, le flux se situe aux alentours de 718 contre 2 305. Enfin, pour un risque de 0,06845%, le flux vaut 1 892,846 contre 8 769,912. Encore une fois, les maximas de blocs se basent sur le quotidien alors que les dépassements de seuil se basent sur toute la période. Dès lors, nous nous attendons à avoir des flux plus élevés pour les maxima de blocs que pour les dépassements de seuil.

Pour conclure cette section, après évaluation des deux adresses IP d'intérêt, nous pouvons constater que la méthode des dépassements de seuil est efficace pour ces deux adresses IP. Par conséquent, cette technique s'avère finalement être valide peu importe la quantité de flux prise en considération.

Enfin, entre les deux méthodes proposées, malgré une équivalence dans les résultats, la méthode des dépassements de seuil semble mieux ajuster les valeurs extrêmes que la méthode des maxima de blocs. Cela pourrait s'expliquer par le fait que la méthode des maxima de blocs mesure ses données de manière quotidienne alors que la méthode des dépassements de seuil mesure ses données de manière globale. Une mesure sur la période entière semble en effet plus réaliste que de manière quotidienne. Cela nous donne dès lors assez confiance en cette méthode des dépassements de seuil pour interpréter ses valeurs.

6. Conclusion

Le trafic Internet est un élément de plus en plus crucial pour les entreprises dans la gestion d'un grand nombre d'aspects importants de leur activité. Par conséquent, il devient essentiel de savoir le gérer. Néanmoins, aucune étude ne fournit une méthode concrète pour permettre aux entreprises de tirer le meilleur parti de leurs données. Ce travail fournit donc aux entreprises une méthode en vue de gérer au mieux la gestion de leur site Internet et/ou de leurs systèmes. Cette méthode utilise la théorie des valeurs extrêmes empruntées aux statistiques. Pour ce faire, deux principes fondamentaux de cette théorie ont été explicités et utilisés, à savoir le principe des maxima de blocs et le principe des dépassements de seuil. Ceux-ci diffèrent l'un de l'autre dans la manière dont ils sélectionnent les données considérées comme extrêmes. La méthode explicitée dans ce travail permet aux entreprises deux choses essentielles. Étant donné un certain flux, quelle est la probabilité de le dépasser? Étant donné un certain niveau de risque, quel est le flux auquel nous pouvons nous attendre à atteindre et à dépasser? Les entreprises pourront dès lors, soit déterminer quand un flux plus extrême que ce qu'elles ont déjà observé apparaîtra, soit déterminer le flux qu'elles devront être capables de gérer en fonction du risque accepté.

Dès lors, cette méthode contribue de manière certaine à la littérature. En effet, une telle technique n'existait pas encore et constitue un vrai plus pour les entreprises. Ces dernières pourront gérer au mieux leur site Internet et/ou leurs systèmes et décider si elles désirent augmenter, diminuer ou ne rien changer en ce qui concerne la capacité de leur site Internet et/ou leurs systèmes. Leur allocation de budget s'en verra améliorée.

Cependant, la limite de ce travail réside dans la taille assez réduite du jeu de données. En pratique, les bases de données des entreprises sont très larges. Dès lors, nous pouvons nous demander ce que cela donnerait si une entreprise avec une grande base de données voulait utiliser la méthode proposée. Est-ce que la méthode peut s'adapter à autant de données ou est-elle justement trop simple pour une base de données de ce type?

Plusieurs perspectives futures sont envisageables pour ce travail. D'abord, l'estimation des paramètres peut être améliorée par l'utilisation de la technique 'profile likelihood'. Ensuite, nous avons remarqué qu'il semblerait y avoir différents niveaux de stabilité présents dans les données, comme plusieurs modèles ou distributions possibles pour les extrêmes modérés et les extrêmes très forts. Cela voudrait dire que ces différents types d'extrêmes ne se comportent pas de la même manière. L'irrégularité est un phénomène que nous ne voyons pas souvent dans les vraies données. L'investigation du flux constitue donc une autre piste. Enfin, l'intégration du concept de dépendance temporelle constitue une autre perspective. Pour l'instant, nous avons supposé l'indépendance des données. Mais si un événement de grande envergure a lieu et que tout le monde veut accéder au site Internet, le flux sera élevé pendant un certain temps, démontrant ainsi une dépendance temporelle très forte. Une amélioration de cette méthode de base serait donc de définir les paramètres comme étant égaux à des fonctions pouvant varier dans le temps.

7. Bibliographie

2020 Progress Update : The 10G Platform. (2020, 3 Janvier). NCTA — The Internet & Television Association. Consulté le 31 Août 2020, à l'adresse https://www.ncta.com/whats-new/2020-progress-update-the-10g-platform

Alasmar, M., Parisis, G., Clegg, R., & Zakhleniu, N. (2019, April). On the Distribution of Traffic Volumes in the Internet and its Implications. In *IEEE INFOCOM 2019-IEEE conference on computer communications* (pp. 955-963). IEEE.

Alasmar, M., & Zakhleniuk, N. (2017). Network link dimensioning based on statistical analysis and modeling of real internet traffic. *arXiv preprint arXiv:1710.00420*.

Ali, K. A. K. (2019, 16 Avril). *SD, HD, 4K, 8K. . . comprendre les définitions des TV, PC et smartphones en 5mn*. CNET France. Consulté le 22 Janvier 2021, à l'adresse https://www.cnetfrance.fr/produits/sd-hd-ultra-hd-4k-8k-comprendre-les-definitions-des-tv-39786402.htm

Bathelot, B. (2011, 2 Décembre). *Trafic web - Définitions Marketing*. Définitions Marketing. Consulté le 16 Mai 2021, à l'adresse https://www.definitions-marketing.com/definition/trafic-web/

Belwaer, O. (2019, 27 Mai). *L'importance du trafic Internet pour le développement de votre activité*. WAOO (Worldwide Agency for Offshore Outsourcing). Consulté le 31 Août 2020, à l'adresse https://www.waoo-digital.com/limportance-du-trafic-internet-pour-le-developpement-de-votre-activite/

Callado, A., Kamienski, C., Szabó, G., Gero, B. P., Kelner, J., Fernandes, S., & Sadok, D. (2009). A survey on internet traffic identification. *IEEE communications surveys & tutorials*, *11*(3), 37-52.

Coles, S. (2001). An introduction to statistical modeling of extreme values. London: Springer-Verlag. ISBN: 1-85233-459-2

Dainotti, A., Pescape, A., & Claffy, K. C. (2012). Issues and future directions in traffic classification. *IEEE network*, *26*(1), 35-40.

Enisa. (s. d.). *Botnets*. Enisa - European Union Agency for Cybersecurity. Consulté le 07 Mai 2021, à l'adresse https://www.enisa.europa.eu/topics/csirts-in-europe/glossary/botnets

Färber, J., Bodamer, S., & Charzinski, J. (1998, September). Measurement and modelling of Internet traffic at access networks. In *Proceedings of the EUNICE* (Vol. 98, pp. 196-203).

How Internet Traffic Changed During the Pandemic. (2020, 1 Juillet). NCTA — The Internet & Television Association. Consulté le 31 Août 2020, à l'adresse https://www.ncta.com/whats-new/how-internet-traffic-changed-during-the-pandemic

Internet traffic. (2021, 29 Mars). Wikipedia. Consulté le 31 Août 2020, à l'adresse https://en.wikipedia.org/wiki/Internet_traffic

Internet Traffic Classification. (2019, 12 Novembre). CAIDA. Consulté le 16 Mai 2021, à l'adresse https://www.caida.org/archive/classification-overview/

Libotte, A. (2018, 3 Octobre). *Netflix est responsable de 15% du trafic Internet mondial*. RTBF Tendance. Consulté le 21 Janvier 2021, à l'adresse https://www.rtbf.be/tendance/techno/detail_netflix-est-responsable-de-15-du-trafic-internet-mondial?id=10035286

Moore, A. W., & Zuev, D. (2005, June). Internet traffic classification using bayesian analysis techniques. In *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems* (pp. 50-60).

Paquet (réseau) - Définition et Explications. (s. d.). Techno-Science.net. Consulté le 16 Mai 2021, à l'adresse https://www.techno-science.net/definition/11437.html

Patard, A. (2020, 4 Février). *Étude sur l'usage d'Internet et des réseaux sociaux dans le monde en 2020*. BDM (Blog du Modérateur). Consulté le 21 Janvier 2021, à l'adresse https://www.blogdumoderateur.com/internet-reseaux-sociaux-2020/

Patard, A. (2020, 27 Avril). *Étude sur l'usage d'Internet, des réseaux sociaux et du mobile au 1er trimestre 2020*. BDM. Consulté le 21 Janvier 2021, à l'adresse https://www.blogdumoderateur.com/internet-reseaux-sociaux-mobile-t1-2020/

Peng, C., Xu, M., Xu, S., & Hu, T. (2017). Modeling and predicting extreme cyber attack rates via marked point processes. *Journal of Applied Statistics*, *44*(14), 2534-2563.

Report : Where Does the Majority of Internet Traffic Come From? (2019, 17 Octobre). NCTA — The Internet & Television Association. Consulté le 31 Août 2020, à l'adresse https://www.ncta.com/whats-new/report-where-does-the-majority-of-internet-traffic-come

Statista. (2019, 25 Mars). *Temps passé par jour à utiliser internet dans le monde en janvier 2019, par pays (en minutes)*. Statista. Consulté le 20 Janvier 2021, à l'adresse https://fr.statista.com/statistiques/985210/temps-passe-sur-internet-par-jour-par-pays-monde/

Trafic Internet. (2020, 15 Août). Wikipédia. Consulté le 31 Août 2020, à l'adresse https://fr.wikipedia.org/wiki/Trafic_Internet

Tsourti, Z., & Panaretos, J. (2004). Extreme-value analysis of teletraffic data. *Computational statistics & data analysis*, *45*(1), 85-103.

Uchida, M. (2004). Traffic data analysis based on extreme value theory and its applications to predicting unknown serious deterioration. *IEICE transactions on information and systems*, *87*(12), 2654-2664.

Utilisateurs d'Internet (% de la population) (s. d.). La Banque Mondiale. Consulté le 21 Janvier 2021, à l'adresse https://donnees.banquemondiale.org/indicator/IT.NET.USER.ZS?end=2019&name_desc=false&start=1990&type=shaded&view=chart&year=2019

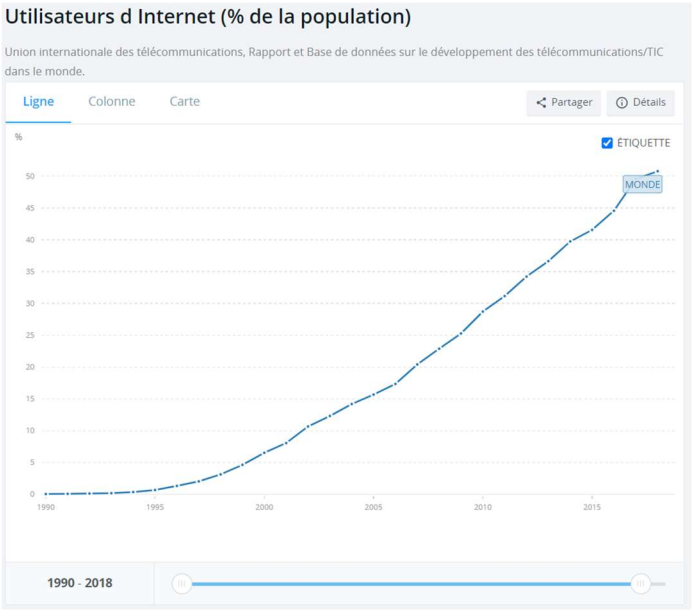
Williamson, C. (2001). Internet traffic measurement. *IEEE internet computing*, *5*(6), 70-74.

Zareen, K. T. (2019, 23 Juillet). *Bandwidth required for HD FHD 4K video streaming*. SYNOPI. Consulté le 22 Janvier 2021, à l'adresse https://www.synopi.com/bandwidth-required-for-hd-fhd-4k-video/

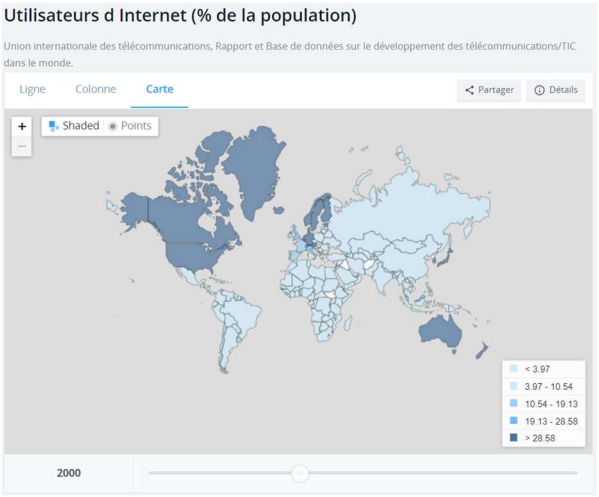
8. Annexes

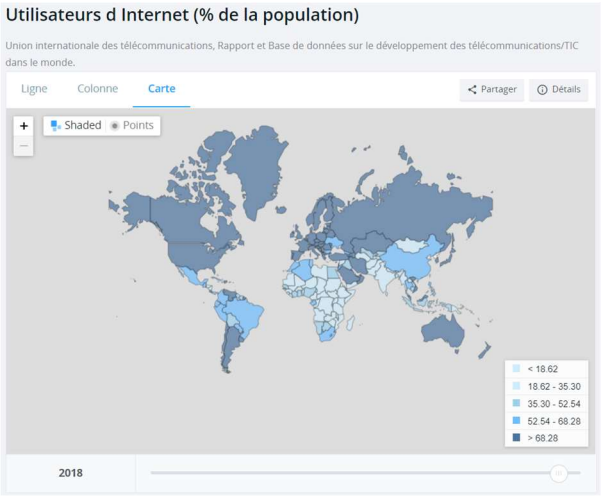
Annexe A – Figures Supplémentaires

Annexe A.1



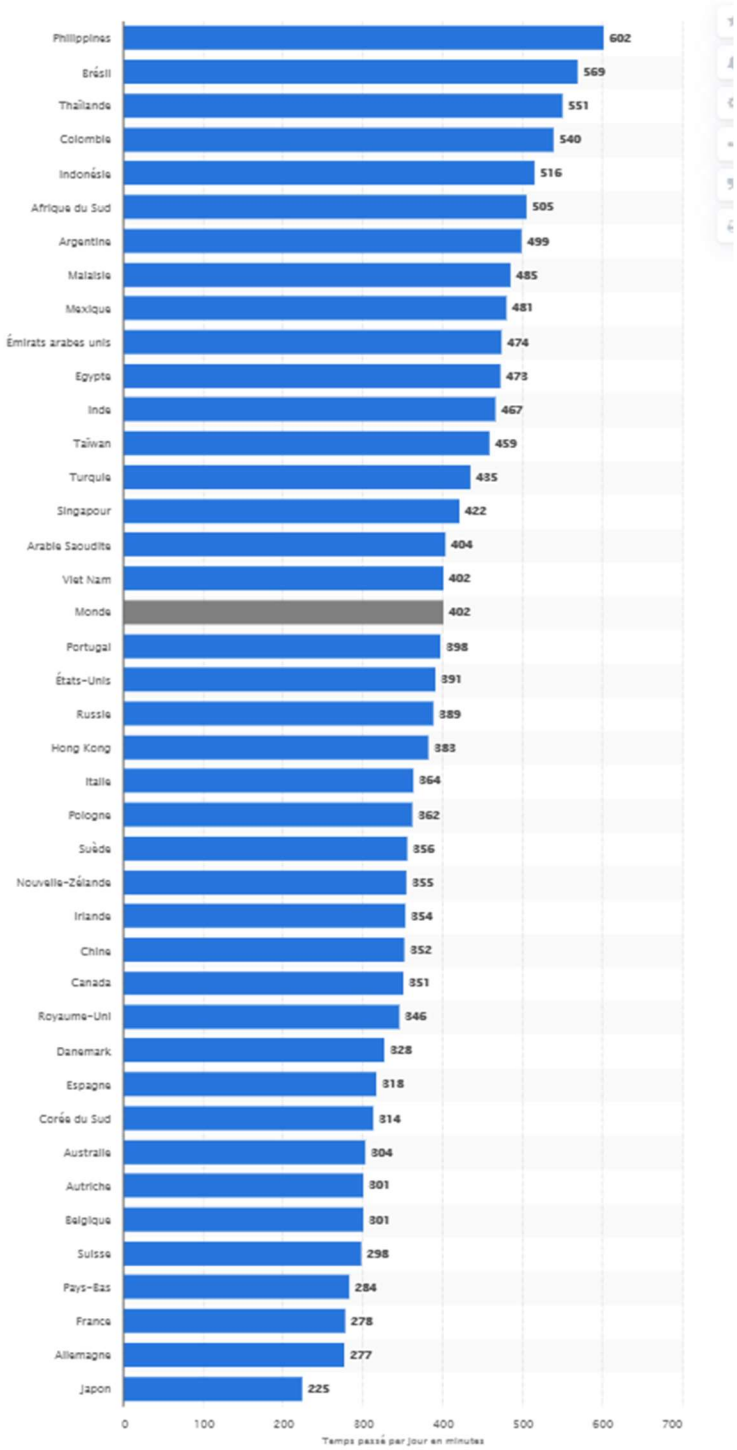
Annexe A.2





→ Source: *Utilisateurs d'Internet (% de la population)* (s. d.). La Banque Mondiale. Consulté le 21 Janvier 2021, à l'adresse https://donnees.banquemondiale.org/indicator/IT.NET.USER.ZS?end=2019&name_desc=false&start=1990&type=shaded&view=chart&year=2019

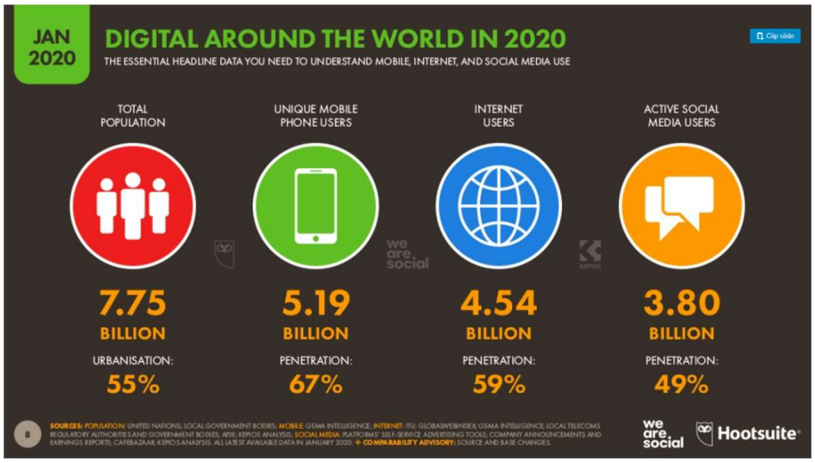
Annexe B



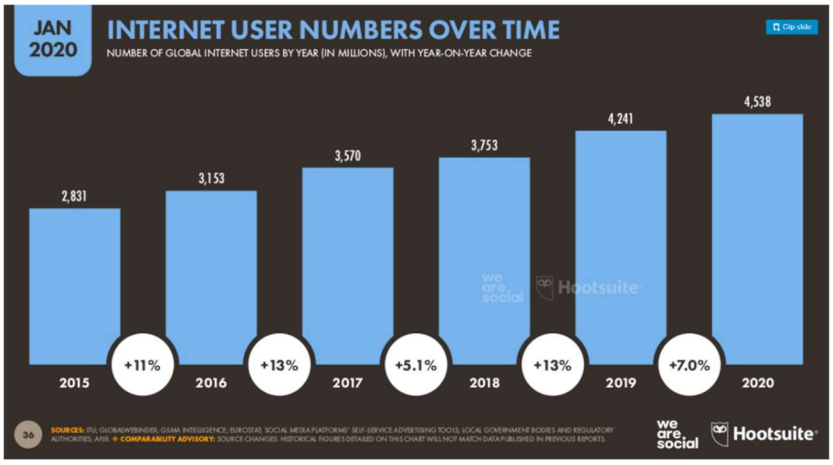
→ Source: Statista. (2019, 25 Mars). *Temps passé par jour à utiliser internet dans le monde en janvier 2019, par pays (en minutes)*. Statista. Consulté le 20 Janvier 2021, à l'adresse https://fr.statista.com/statistiques/985210/temps-passe-sur-internet-par-jour-par-pays-monde/

Annexe C – Figures Supplémentaires

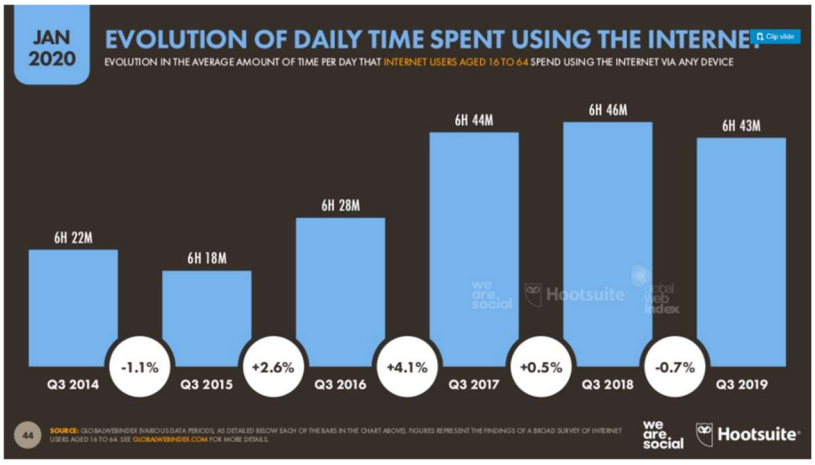
Annexe C.1



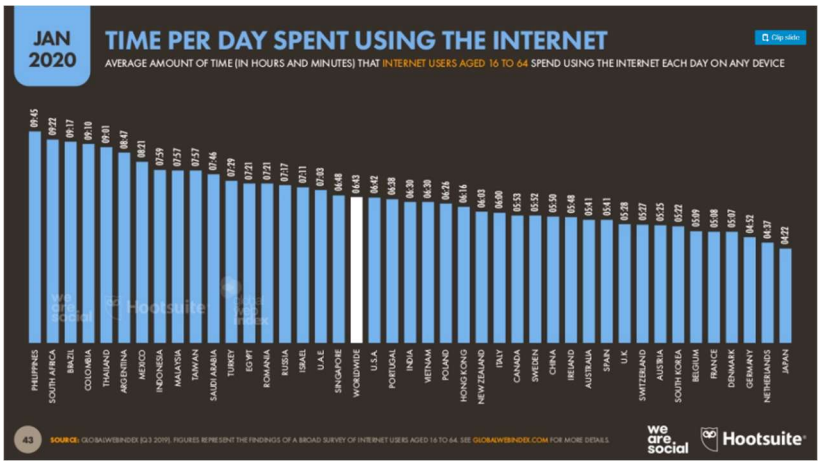
Annexe C.2



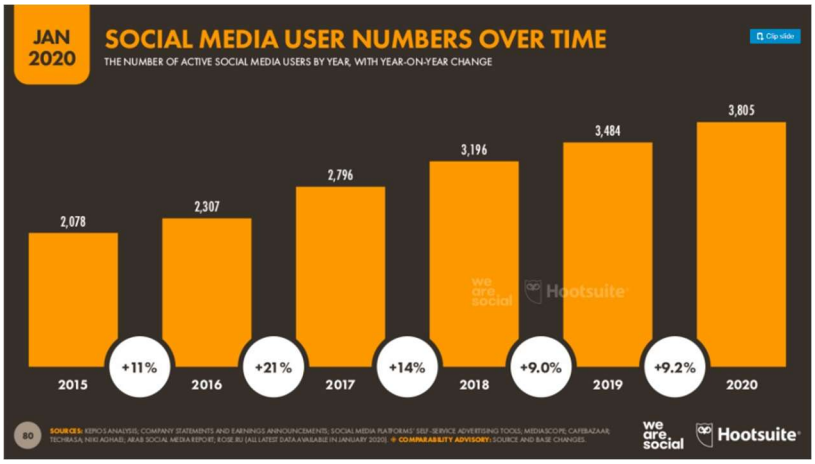
Annexe C.3



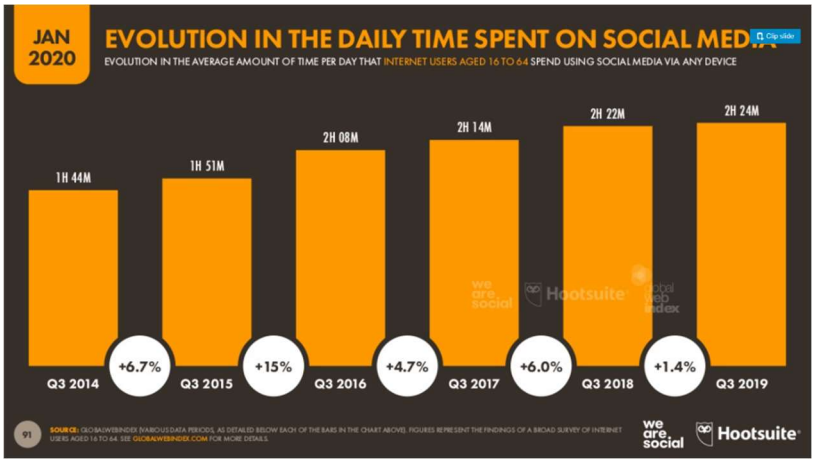
Annexe C.4



Annexe C.5



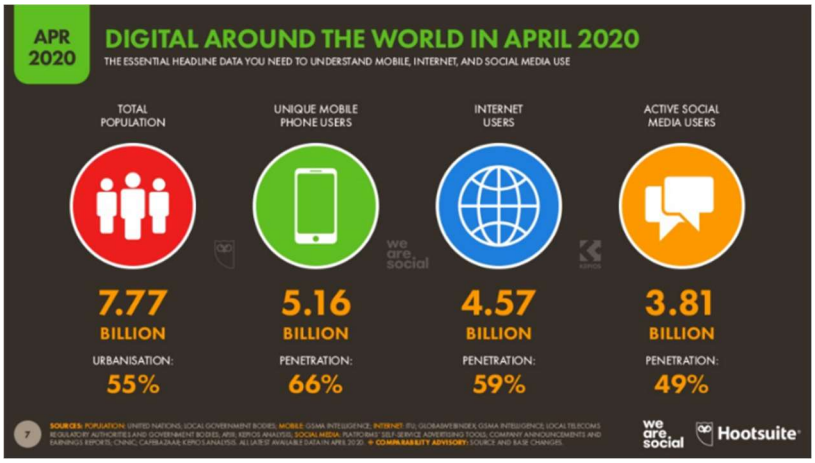
Annexe C.6



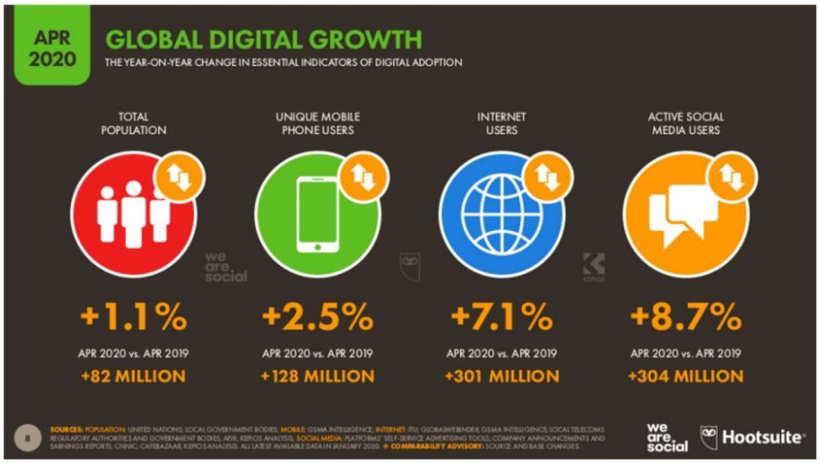
→ Source: Patard, A. (2020, 4 Février). *Étude sur l'usage d'Internet et des réseaux sociaux dans le monde en 2020*. BDM (Blog du Modérateur). Consulté le 21 Janvier 2021, à l'adresse https://www.blogdumoderateur.com/internet-reseaux-sociaux-2020/

Annexe D – Figures Supplémentaires

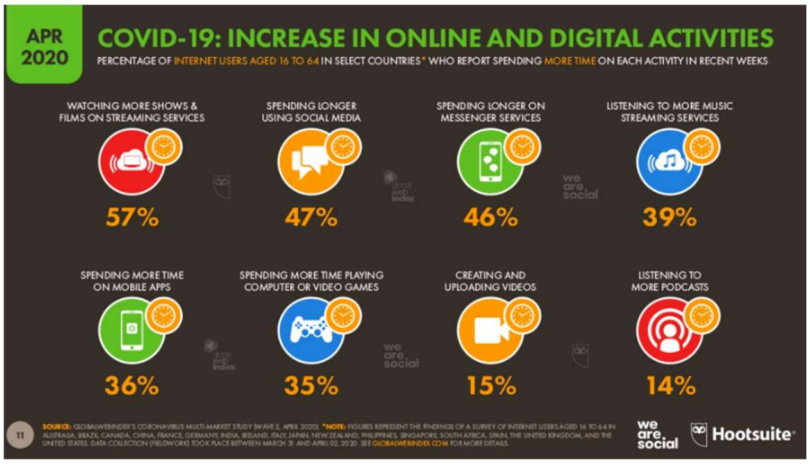
Annexe D.1



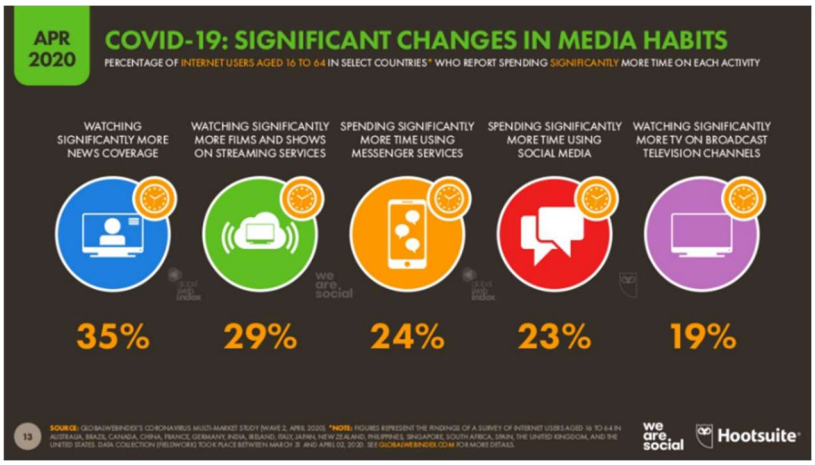
Annexe D.2



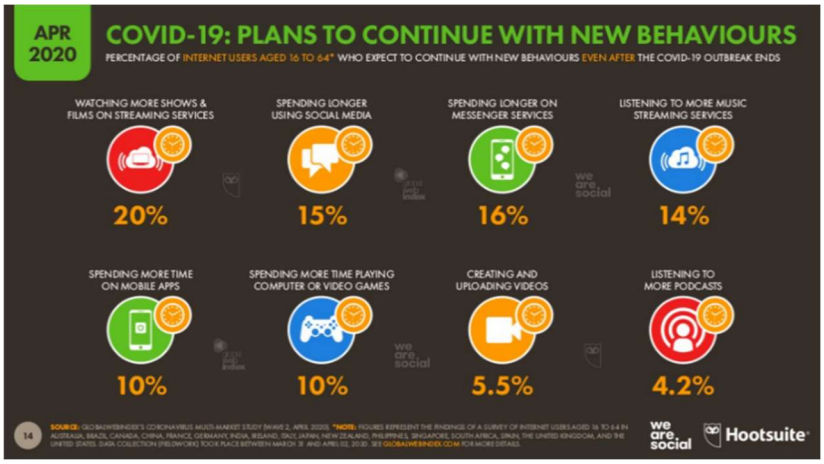
Annexe D.3



Annexe D.4



Annexe D.5



→ Source: Patard, A. (2020, 27 Avril). *Étude sur l'usage d'Internet, des réseaux sociaux et du mobile au 1er trimestre 2020*. BDM. Consulté le 21 Janvier 2021, à l'adresse https://www.blogdumoderateur.com/internet-reseaux-sociaux-mobile-t1-2020/

Annexe E – Figures Supplémentaires

Annexe E.1

Source: Libotte, A. (2018, 3 Octobre). *Netflix est responsable de 15% du trafic Internet mondial*. RTBF Tendance. Consulté le 21 Janvier 2021, à l'adresse https://www.rtbf.be/tendance/techno/detail_netflix-est-responsable-de-15-du-trafic-internet-mondial?id=10035286

Annexe E.2

Source: *Report : Where Does the Majority of Internet Traffic Come From?* (2019, 17 Octobre). NCTA — The Internet & Television Association. Consulté le 31 Août 2020, à l'adresse https://www.ncta.com/whats-new/report-where-does-the-majority-of-internet-traffic-come